REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-06-2011	Final Performance Report	Feb 1, 2008 -- March 31, 2011

4. TITLE AND SUBTITLE
Towards Statistically Undetectable Steganography

5a. CONTRACT NUMBER
FA9550-08-1-0084

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Prof. Jessica Fridrich

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Research Foundation of State University of New York
SUNY at Binghamton
4400 Vestal parkway East
Binghamton, NY 13902-6000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF, AFRL
AF Office of Scientific Research
875 N. Randolph Street RM 3112
Arlington, VA

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR/PKA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0142

12. DISTRIBUTION/AVAILABILITY STATEMENT
All data delivered is approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Fundamental asymptotic laws for imperfect steganography were established. The size of the secure payload was shown to be proportional to the square root of the number of elements (pixels) in the digital media (image); the root rate and the steganographic Fisher information were established as the appropriate measures for evaluating the capacity and for benchmarking steganographic systems in digital media. A general framework for building steganographic systems by minimizing embedding distortion was established by drawing a connection between statistical physics and steganography. The framework, termed the Gibbs construction, allows computing fundamental bounds between distortion and statistical detectability, simulate optimal embedding methods, and construct practical embedding algorithms using syndrome-trellis codes. The framework has essentially narrowed down the design of secure steganographic schemes to the task of designing the distortion function. The PI has also proposed a method for optimizing the distortion function to minimize statistical detectability. Practical merit of the achievements was demonstrated by building new steganographic methods with markedly improved security w.r.t. existing art for images in spatial and JPEG formats.

15. SUBJECT TERMS
Steganography, covert communication, statistical detectability, asymptotic performance, secure payload, minimum-distortion steganography

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Jessica Fridrich
Unclassified	Unclassified	Unclassified	Unlimited		19b. TELEPHONE NUMBER *(Include area code)* 607 777 6177

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

FINAL REPORT FOR CONTRACT FA9550-08-1-0084 ENTITLED "TOWARDS STATISTICALLY UNDETECTABLE STEGANOGRAPHY"

OBJECTIVES

(1) To study the fundamental security limitations of covert communication systems that embed secret data in empirical cover objects, such as digital imagery.
(2) Derive performance bounds of steganographic schemes that embed secret messages by minimizing a distortion function, construct simulators of optimal (bound-reaching) schemes, and design algorithms for practical near-optimal schemes.
(3) Investigate the possibility to develop high-capacity steganographic schemes undetectable by a given steganalysis detector and formulate the implications for improving detection of steganography.

PERSONNEL SUPPORTED

- Prof. Jessica Fridrich, PI
- Dr. Miroslav Goljan, Postdoctoral Assistant
- Tomáš Filler, PhD student
- Jan Kodovský, PhD student
- Tomas Pevný, PhD student

LIST OF PUBLISHED PAPERS AND DISSERTATIONS

Books.

(1) J. Fridrich, Steganography in Digital Media: Principles, Algorithms, and Applications, Cambridge University Press, November 2009.

Journal papers.

(1) T. Filler, J. Judas, and J. Fridrich, "Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes," to appear in IEEE Trans. on Info. Forensics and Security, 2011.
(2) J. Kodovský and J. Fridrich, "Quantitative Structural Steganalysis of Jsteg," IEEE Trans. on Info. Forensics and Security, vol. 5(4), pp. 681–693, 2010.
(3) T. Filler and J. Fridrich, "Gibbs construction in Steganography," IEEE Trans. on Info. Forensics and Security, vol. 5(4), pp. 705–720, 2010.
(4) T. Pevný, A. Ker, and J. Fridrich, "From Blind to Quantitative Steganalysis," to appear in IEEE Trans. on Info. Forensics and Security, 2011.
(5) T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by Subtractive Pixel Adjacency Matrix," IEEE Trans. on Info. Forensics and Security, vol. 5(2), pp. 215–224, 2010.
(6) T. Filler and J. Fridrich, "Fisher Information Determines Capacity of ϵ-secure Steganography," 11th Information Hiding Workshop, Darmstadt, Germany, June 7–10, 2009, LNCS vol. 5806, Springer-Verlag, pp. 31–47.
(7) J. Fridrich, Asymptotic Behavior of the ZZW Embedding Construction, IEEE Trans. on Info. Forensics and Security, vol. 4(1), pp. 151–153, March 2009.
(8) T. Pevný and J. Fridrich, "Benchmarking for Steganography," 10th Information Hiding Workshop, Santa Barbara, California, May 19–21, LNCS, vol. 5284, Springer-Verlag, pp. 251–267, 2008.
(9) T. Pevný and J. Fridrich, "Multi-Class Detector of Current Steganographic Methods for JPEG Format," IEEE Trans. on Info. Forensics and Security, vol. 3(4), pp. 635–650, December 2008.
(10) T. Pevný and J. Fridrich, "Detection of Double-Compression in JPEG Images for Applications in Steganography," IEEE Trans. on Info. Forensics and Security, vol. 3(2), pp. 247–258, 2008.

Conference papers.

(1) T. Filler and J. Fridrich, "Design of Adaptive Steganographic Schemes for Digital Images," Proc. SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics XIII, San Francisco, CA, January 23–26, pp. OF 1–14, 2011.
(2) T. Filler and J. Fridrich, "Minimizing Additive Distortion Functions with Non-Binary Embedding Operation in Steganography," Proc. IEEE WIFS'10, Seattle, WA, December 12–15, 2010.
(3) T. Filler and J. Fridrich, "Steganography Using Gibbs Random Fields," ACM Multimedia & Security Workshop, Rome, Italy, September 9–10, 2010.
(4) J. Kodovský, T. Pevný, and J. Fridrich, "Modern Steganalysis Can Detect YASS," Proc. SPIE, Electronic Imaging, Media Forensics and Security XII, San Jose, CA, January 17–21, pp. 02-01–02-11, 2010.
(5) T. Filler, J. Judas, and J. Fridrich, "Minimizing Embedding Impact in Steganography Using Trellis-Coded Quantization," Proc. SPIE, Electronic Imaging, Media Forensics and Security XII, San Jose, CA, January 17–21, pp. 05-01–05-14, 2010.
(6) T. Filler and J. Fridrich, "Wet ZZW Construction," Proc. IEEE WIFS'09, London, UK, December 6–9, 2009.
(7) J. Kodovský and J. Fridrich, "Calibration Revisited," Proc. ACM Multimedia and Security Workshop, Princeton, NJ, September 7–8, pp. 63–74, 2009.
(8) T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by Subtractive Pixel Adjacency Matrix," Proc. ACM Multimedia and Security Workshop, Princeton, NJ, September 7–8, pp. 75–84, 2009.
(9) T. Filler and J. Fridrich, "Complete Characterization of Perfectly Secure Stegosystems with Mutually Independent Embedding Operation," Proc. IEEE ICASSP, April 19–24, 2009.\
(10) T. Pevný, J. Fridrich, and A. D. Ker, "From Blind to Quantitative Steganalysis," Proc. SPIE, Electronic Imaging, Media Forensics and Security XI, San Jose, CA, January 18–22, pp. 0C 1–0C 14, 2009.
(11) T. Filler, J. Fridrich, and A. D. Ker, "The Square Root Law of Steganographic Capacity for Markov Covers," Proc. SPIE, Electronic Imaging, Media Forensics and Security XI, San Jose, CA, January 18–22, pp. 08 1–08 11, 2009.
(12) A. D. Ker, T. Pevný, J. Kodovský, and J. Fridrich, "The Square Root Law of Steganographic Capacity," Proc. ACM Multimedia and Security Workshop, Oxford, UK, September 22–23, pp. 107–116, 2008.

(13) J. Kodovský, and J. Fridrich, "On Completeness of Feature Spaces in Steganalysis," Proc. ACM Multimedia and Security Workshop, Oxford, UK, September 22–23, pp. 123–132, 2008.
(14) T. Pevný and J. Fridrich, "Novelty Detection in Blind Steganalysis," Proc. ACM Multimedia and Security Workshop, Oxford, UK, September 22–23, pp. 167–176, 2008.
(15) J. Fridrich and T. Pevný, "Estimation of Primary Quantization Matrix for Steganalysis of Double-Compressed JPEG Images," Proc. SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, San Jose, CA, January 26–31, pp. 11-1–11-13, 2008.

All papers were presented at the above noted conferences or published in respective journals. The No Cost Extension (Jan 1 – Mar 31, 2011) gave the PI the opportunity to finish work on conference paper no. 1) and T. Filler's PhD dissertation.

PhD dissertations.

(1) T. Filler, "Imperfect Stegosystems – Asymptotic Laws and Near-Optimal Constructions," Ph.D. Dissertation, SUNY Binghamton, Dept. of ECE, April 2011.
(2) J. Kodovský, "Advances in Feature-Based Steganalysis of Digital media," Ph.D. Dissertation, SUNY Binghamton, Dept. of ECE, in preparation, December 2011.

EXECUTIVE SUMMARY

Steganography is a relatively new mode of communication that is far less developed and understood in comparison to other research fields that offer privacy and security, such as cryptography. In particular, despite its fundamental importance, the very basic question of how big a payload it is secure to embed in a given cover object to prevent an adversary from detecting the presence of a secret message, is still open. The first result in this direction appeared in [61, 88, 89]. Through a series of articles published in 2004–2008, it has been established that steganographic capacity of *perfectly secure* stegosystems grows *linearly* with the number of cover elements (pixels) – secure steganography has a positive rate. In practice, however, neither the adversary (Warden) nor the steganographer has perfect knowledge of the cover source and thus it is unlikely that perfectly secure stegosystems for *empirical* covers, such as digital media, will ever be constructed. This justifies the first topic on which the PI focused – the study of secure capacity of *imperfect* stegosystems.

In 2006 and 2007, theoretical results appeared concerning the paradigm of batch steganography [50, 51] (embedding by dividing payload among multiple covers to minimize statistical detectability). These results, supported by experiments with blind steganalyzers, pointed to an emerging paradigm: whether steganography is performed in a large batch of cover objects or in a single large object, the size of the secure payload grows only sub-linearly. In particular, the secure payload appeared to be proportional to the *square root* of the number of pixels in the cover image. This so-called Square-Root Law of Steganography is the first principal achievement reported here. In Section 1, it is formally established for imperfect stegosystems that hide messages in covers modeled as a stationary Markov chain and when the embedding changes are mutually independent. An important new part of this contribution, explained in Section (2), is a complete characterization of perfectly-secure cover sources with respect to a given embedding operation. The square root law has been confirmed experimentally on real imagery and several different embedding and steganalysis methods in Section 3. Among other contributions, it has been established that secure payload of imperfect steganographic systems is completely described using the so-called *root rate* that depends on the steganographic Fisher information, which forms a security descriptor equivalent to the Kullback–Leibler divergence (Section 2). The Fisher information can be used for optimizing the design of steganographic systems, for benchmarking, and comparison of covert communications schemes.

Having established the fundamental limitations of covert communication in empirical (incognizable) covers, the PI then directed her effort towards a general framework within which steganographic schemes can be built in practice (Sections 5–6). By abandoning the concept of preserving the cover distribution (as the distribution is incognizable anyway), the steganography is instead casted as a source coding with fidelity constraint. The so-called minimal embedding impact steganography is formulated using the concept of distortion that is fundamentally tied to statistical detectability. The PI resolved the problem of embedding while minimizing an essentially arbitrary distortion function by formulating the embedding problem within statistical physics. The resulting "Gibbs construction" (Section 6) is an elegant framework within which one can study, design, and optimize steganographic schemes. Syndrome–trellis codes were proposed as a general approach to practical embedding. Finally, the PI managed to tie distortion to statistical detectability by converting the problem of minimizing detection to a parameter optimization problem [24]. Thus, the developed framework provides an enclosed and complete theoretical basis for further development of steganography in empirical covers.

This report makes use of the Iverson bracket $[I]$ defined to be 1 if the logical expression I is true and zero otherwise. The binary entropy function $h(x) = -x \log x - (1-x)\log(1-x)$ is expressed in bits.

1. The Square Root Law of Secure Steganographic Payload

In steganography, the sender communicates with the receiver by hiding her messages inside innocuous looking (cover) objects. Most practical steganographic methods embed messages by slightly modifying individual elements of the cover, obtaining thus the modified stego object that conveys the hidden message. The goal here is to make the stego objects statistically indistinguishable from covers – a passive warden, who is merely inspecting the traffic, cannot construct a detector of stego objects that would work better than an algorithm that makes random guesses. The assumption is that, up to a secret shared key, the warden is familiar with all details of the steganographic scheme: this is the so-called Kerckhoffs' principle, which is also interpreted to mean that the warden has complete knowledge of the probabilistic distribution of cover objects.

Statistical detectability of embedding changes depends on their character and extent. Intuitively, it should be possible to send short messages with lower risk of being detected than long messages. From a practical point of view, the sender needs to know how long a message she can embed for a chosen risk – she needs to know the steganographic capacity of the stegosystem. Unfortunately, determining the steganographic capacity analytically for real digital media objects, such as digital images, is very difficult even for the simplest steganographic paradigms, such as LSB (Least Significant Bit) embedding. The reason is the lack of accurate statistical models.

One may intuitively expect the steganographic capacity to be linear in the size of the cover object by referring to a similar result for capacity of noisy communication channels. This is, indeed, valid if the stegosystem is perfectly secure, since there is no possible detector [61, 12]. In view of the absence of provably secure steganographic methods for real digital media, it makes sense to investigate steganographic capacity of imperfect embedding methods for which detectors exist and inquire about the largest payload that can be embedded using their ϵ-secure versions in the sense of Cachin [9].

The fact that steganographic capacity is most likely sublinear was already suspected by Anderson [1] in 1996:

> "Thanks to the Central Limit Theorem, the more covertext one gives the Warden, the better he will be able to estimate its statistics, and so the smaller the rate at which [the steganographer] will be able to tweak bits safely. The rate might even tend to zero..."

The analysis of batch steganography and pooled steganalysis by Ker [51] tells us that steganographic capacity of imperfect stegosystems only grows as the square root of the number of communicated covers. This result could be interpreted as the square root capacity law for a single image by dividing it into smaller blocks. The capacity result, however, was obtained with the assumption that the individual images (blocks) form a sequence of *independent* random variables, which is clearly false not only for images but also other digital media files. The main contribution reported in this section is to establish the same law for the simplest form of dependence that enables analytical reasoning – it will be assumed that individual elements of the cover (e.g., pixels) follow stationary Markov chain.

1.1. Basic assumptions.

This section introduces notation and three basic assumptions under which the SRL (Square-Root Law) is proved. The first assumption concerns the impact of embedding. It is postulated that the stego object is obtained by applying a mutually independent embedding operation to each cover element. This type of embedding can be found in majority of practical embedding methods (see, e.g., [37] and the references therein). The second assumption concerns the model of covers. The individual cover elements are required to form a first-order Markov chain because this model is analytically tractable while allowing study of more realistic cover sources with memory. Finally, the third assumption essentially states that the steganographic method is not perfectly secure.

Throughout the report, $\mathbb{A} = (a_{ij})$ denotes a matrix with elements a_{ij}, calligraphic font ($\mathcal{X}$) to denote sets, and capital letters (X, Y) to denote random variables, both vector and scalar. If y is a vector with components $y = (y_1, \ldots, y_n)$, y_k^l denotes the subsequence $y_k^l = (y_k, \ldots, y_l)$. If $Y = (Y_1, \ldots, Y_n)$ is a random vector with underlying probability distribution P, then $P(Y_k^l = y_k^l)$ denotes the marginal probability $P(Y_k = y_k, Y_{k+1} = y_{k+1}, \ldots, Y_l = y_l)$.

An n-element cover source will be represented using a random variable $X_1^n \triangleq (X_1, \ldots, X_n)$ distributed according to some general distribution $P^{(n)}$ over $\mathcal{X}^n$, $\mathcal{X} \triangleq \{1, \ldots, N\}$. A specific cover object is a realization of X_1^n and will be denoted with the corresponding lower case letter $x_1^n \triangleq (x_1, \ldots, x_n) \in \mathcal{X}^n$. A stegosystem, with covers of fixed size n, is a triple $S_n = (X_1^n, \Phi^{(n)}, \Psi^{(n)})$ consisting of the random variable describing the cover source, embedding mapping $\Phi^{(n)}$, and extraction mapping $\Psi^{(n)}$. The embedding mapping $\Phi^{(n)}$ applied to X_1^n induces another random variable $Y_1^n \triangleq (Y_1, \ldots, Y_n)$ with probability distribution $Q_\beta^{(n)}$ over $\mathcal{X}^n$. Specific realizations of Y_1^n are called stego objects and will be denoted $y_1^n \triangleq (y_1, \ldots, y_n)$. Here, $\beta \geq 0$ is a scalar parameter of embedding whose meaning will be explained shortly. The specific details of the embedding (and extraction) mappings are immaterial for this study. In particular, one only needs to postulate the probabilistic *impact* of embedding.

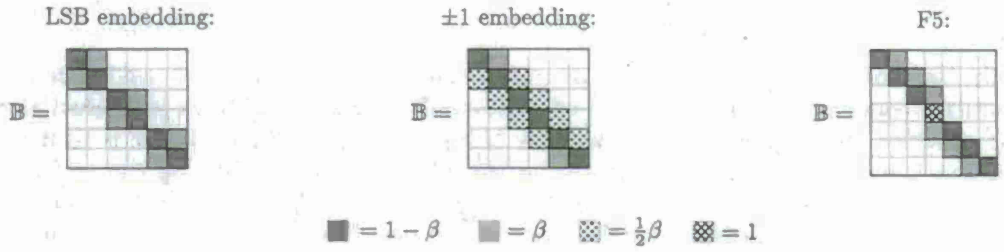

FIGURE 1.1. Examples of several embedding methods in the form of a functional matrix $\mathbb{B}$.

Assumption 1: [Mutually independent embedding] The embedding algorithm visits every cover element X_k and, independently of all other elements, modifies it to a corresponding element of the stego object Y_k with probability

$$Q_\beta(Y_k = j|X_k = i) \triangleq b_{i,j}(\beta) = \begin{cases} 1+\beta c_{i,i} & \text{if } i = j \\ \beta c_{i,j} & \text{otherwise,} \end{cases} \tag{1.1}$$

for some constants $(c_{i,j})$ with $c_{i,j} \geq 0$ for $i \neq j$. Note that because $\sum_{j\in\mathcal{X}} b_{i,j} = 1$, one must have $c_{i,i} = -\sum_{j\neq i} c_{i,j}$ for each $i \in \mathcal{X}$. The matrix $(c_{i,j})$ reflects the inner workings of the embedding algorithm, while the parameter β captures the *extent* of embedding changes. It will be useful to think of β as the relative number of changes (change rate) or some function of the change rate. Also note that one can find sufficiently small β_0, such that $b_{i,i}(\beta) > 0$ for $\beta \in [0, \beta_0]$ and all $i \in \mathcal{X}$.

Because the matrix $\mathbb{B}_\beta \triangleq (b_{i,j}(\beta))$ does not depend on $k \in \{1, \ldots, n\}$ or the history of embedding changes, one can say that the stego object is obtained from the cover by applying to each cover element a Mutually Independent embedding operation (one speaks of *MI embedding*). The independence of embedding modifications implies that the conditional probability of stego object given the cover object can be factorized, i.e., $Q_\beta^{(n)}(Y_1^n|X_1^n) = \prod_{i=1}^n Q_\beta(Y_i|X_i)$.

Many embedding algorithms across different domains use MI embedding. Representative examples are LSB embedding, ±1 embedding, stochastic modulation, Jsteg, MMx, and various versions of the F5 algorithm [29]. Examples of the matrix $\mathbb{B}_\beta$ for three selected embedding methods are shown in Figure 1.1.

Next, an assumption on the cover source is formulated.

Assumption 2: [Markov cover source] It is assumed that the cover source X_1^n is a first-order stationary Markov Chain over $\mathcal{X}$, which will often be abbreviated as simply Markov Chain (MC). This source is completely described by its stochastic transition probability matrix $\mathbb{A} \triangleq (a_{ij}) \in \mathbb{R}^{N\times N}$, $a_{ij} = Pr(X_k = j|X_{k-1} = i)$, and by the initial distribution $Pr(X_1)$. The probability distribution induced by the MC source generating n-element cover objects satisfies $P^{(n)}(X_1^n = x_1^n) = P^{(n-1)}(X_1^{n-1} = x_1^{n-1})a_{x_{n-1}x_n}$, where $P^{(1)}(X_1)$ is the initial distribution. Furthermore, it is assumed that the transition probability matrix of the cover source satisfies $a_{ij} \geq \delta > 0$, for some δ and thus the MC is irreducible. The stationary distribution of the MC source is a vector $\pi \triangleq (\pi_1, \ldots, \pi_N)$ satisfying $\pi\mathbb{A} = \pi$. In this report, it will always be assumed that the initial distribution $P^{(1)}(X_1) = \pi$, which implies $P^{(n)}(X_k) = \pi$ for every n and k. This assumption simplifies the analysis without loss of generality because the marginal probabilities $P^{(n)}(X_k)$ converge to π with exponential rate w.r.t. k (see Doob [17], equation (2.2) on page 173). In other words, MCs are "forgetting" their initial distribution with exponential rate.

Under the above assumption and the class of MI embedding, the source of stego objects no longer exhibits the Markov property and forms a Hidden Markov Chain (HMC) instead [79]. The HMC model is described by its hidden states (cover elements) and output transition probabilities (MI embedding). Hidden states are described by the cover MC and the output probability transition matrix $\mathbb{B}$ is taken from the definition of MI embedding.

Unless stated otherwise, in the rest of this report $Q_\beta^{(n)}$ denotes the probability measure induced by the HMC source embedded with parameter β into n-element MC cover objects. By the stationarity of the MC source, the marginal probabilities $P^{(n)}(X_k^{k+1}) = P^{(2)}(X_1^2)$ and $Q_\beta^{(n)}(Y_k^{k+1}) = Q_\beta^{(n)}(Y_1^2)$ for all k. Sometimes the number of elements, n, will be omitted and P and Q_β will denote the probability distribution over cover and stego objects, respectively.

The third assumption concerns the entire stegosystem S_n. Because it is known [61, 12] that steganographic capacity of perfectly secure stegosystems is linear in n, the SRL can only apply to imperfect stegosystems.

Assumption 3: [FI condition] It is assumed that the stegosystem $S_n = (X_1^n, \Phi^{(n)}, \Psi^{(n)})$ is not perfectly secure in the sense of Cachin [9] (the KL divergence $D_{KL}(P^{(n)}||Q_\beta^{(n)}) > 0$). It is shown in Section (2), that for the special case of Markov cover sources X_1^n and MI embedding $\Phi^{(n)}$, this assumption can be equivalently stated in two different forms:

(1) The pair $(P^{(2)}, Q_\beta^{(2)})$ *does not* satisfy so called *Fisher Information condition,*

$$\forall y_1^2 \in \mathcal{X}^2 \quad \left(P^{(2)}(X_1^2 = y_1^2) > 0\right) \;\Rightarrow\; \left(\frac{d}{d\beta} Q_\beta^{(2)}(y_1^2)\big|_{\beta=0} = 0\right). \tag{1.2}$$

(2) There exists a pair of states (i, j) such that

$$P(X_1^2 = (i,j)) \neq Q_\beta(Y_1^2 = (i,j)) \text{ for all } \beta{>}0. \tag{1.3}$$

For the sake of continuity, the PI only provides a few brief arguments. First of all, perfectly secure stegosystems must satisfy (1.2) because the Fisher information

$$I(0) = E_P\left[\left(\frac{d}{d\beta} Q_\beta^{(2)}(y_1^2)\big|_{\beta=0}\right)^2\right]$$

appears as a coefficient in front of β^2 in the Taylor expansion of KL divergence $D_{KL}(P^{(2)}||Q_\beta^{(2)})$ w.r.t. β and thus $\frac{d}{d\beta} Q_\beta^{(2)}(y_1^2)\big|_{\beta=0}$ must be zero whenever $P^{(2)}(y_1^2) > 0$. The opposite implication (zero Fisher information implies zero KL divergence) is not valid in general but holds for MI embedding as shown in Section (2). The second condition follows from the fact that second-order marginal statistics fully describe the first-order MC process and thus if (1.3) does not hold, then both cover and stego distributions are the same for all n (the stegosystem is perfectly secure).

Finally, it should be stressed that Assumptions 1–3 are not overly restrictive and will likely be satisfied for all practical steganographic schemes in some appropriate representation of the cover. For example, in digital images it is unlikely that the distribution of each pixel depends only on its neighbor, but the dependency is likely to be spatially-limited. Then the image can be modeled as a Markov chain made up of overlapping pixel groups. Furthermore, if a stegosystem preserves the first-order statistics of a cover source, it is likely to be detectable by considering higher-order dependencies: the apparently-perfect stegosystem becomes imperfect when the cover is represented by pairs or groups of pixels, coefficients, or some other derived quantities.

1.2. **The square root law of steganographic capacity.** This section contains the formulation and the proof of the main result, which states that the steganographic capacity of imperfect stegosystems with Markov covers and mutually independent embedding operation only grows with the square root of the number of cover elements. This finding has some fundamental implications in steganography and steganalysis. Probably the most remarkable one is that steganographic capacity exhibits quite different properties when compared with capacity of noisy channels or lossless compression. For example, while a mismatch in source model decreases the compression gain by a constant (the KL divergence between the source model and true source distribution), a cover model mismatch in steganography leads to vanishing capacity.

The steganographer is *at risk* (w.r.t. some fixed tuple (P_{FA}^*, P_{MD}^*), with $0 < P_{FA}^* < 1$ and $0 < P_{MD}^* < 1 - P_{FA}^*$) if the warden has a detector with probability of false alarms and missed detection P_{FA}, P_{MD} satisfying $P_{FA} < P_{FA}^*$ and $P_{MD} < P_{MD}^*$.

Theorem 1: [The square root law of steganography for Markov covers] *For a sequence of stegosystems $(S_n)_{n=1}^\infty$ satisfying Assumptions 1–3, the following holds:*

(1) *If the sequence of embedding parameters $\beta(n)$ increases faster than $1/\sqrt{n}$ in the sense that $\lim_{n\to\infty} \frac{\beta(n)}{1/\sqrt{n}} = \infty$, then, for sufficiently large n, the Steganographer is at risk for arbitrary tuple (P_{FA}^*, P_{MD}^*).*

(2) *If $\beta(n)$ increases slower than $1/\sqrt{n}$, $\lim_{n\to\infty} \frac{\beta(n)}{1/\sqrt{n}} = 0$, then the stegosystem can be made ε-secure for any $\varepsilon > 0$ for sufficiently large n. This implies that the Steganographer is not at risk, for any tuple (P_{FA}^*, P_{MD}^*).*

(3) *Finally, if $\beta(n)$ grows asymptotically as fast as $1/\sqrt{n}$, $\lim_{n\to\infty} \frac{\beta(n)}{1/\sqrt{n}} = \epsilon$ for some $0 < \epsilon < \infty$, then the stegosystem is asymptotically $C\epsilon^2$-secure for some constant C.*

Proof: Each part of the theorem is proved separately. From the Kerckhoffs' principle, the warden knows the distribution of cover objects $P^{(n)} = Q_0^{(n)}$.

Part 1 [Steganographer at risk] Here, one needs to prove that the Steganographer is at risk w.r.t. any (P_{FA}^*, P_{MD}^*) for all sufficiently large n. This means that a sequence of detectors, D_n, needs to be constructed for the following composite

binary hypothesis testing problem

$$\begin{aligned} \mathrm{H}_0 &: \beta = 0 \\ \mathrm{H}_1 &: \beta > 0 \end{aligned}$$

based on observing one stego object (one realization of a random sequence with distribution $Q_\beta^{(n)}$). The error probabilities of these detectors are required to satisfy $P_{FA} < P_{FA}^*$ and $P_{MD} < P_{MD}^*$ for all sufficiently large n. The test statistic for each detector D_n is described next.

Equation (1.3) in Assumption 3 guarantees the existence of pair of states (i,j) such that $P(X_1^2 = (i,j)) \neq Q_\beta(Y_1^2 = (i,j))$ for all $\beta > 0$. Thus, the test statistic $\nu_{\beta,n}$ for detector D_n is defined as

$$\nu_{\beta,n} = \sqrt{n}\left|\frac{1}{n-1}h_\beta[i,j] - P\big(X_1^2 = (i,j)\big)\right|, \tag{1.4}$$

where $\frac{1}{n-1}h_\beta[i,j]$ is the relative count of the number of consecutive pairs (i,j) in an n-element stego object embedded using parameter β (In terms of indicator functions[1], $h_\beta[i,j] = \sum_{k=1}^{n-1} \mathbb{I}_{\{Y_k=i, Y_{k+1}=j\}}$). Note that due to stationarity of the cover source, $E\left[\frac{1}{n-1}h_\beta[i,j]\right] = Q_\beta(Y_1^2 = (i,j))$ for all β.

The following is proved for the difference between the means of $\nu_{\beta,n}$ under both hypotheses:

$$\lim_{n\to\infty} E[\nu_{\beta,n}] - E[\nu_{0,n}] = \infty \text{ when } \sqrt{n}\beta \to \infty. \tag{1.5}$$

Suppose, for a contradiction, that there exists $K > 0$, and a strictly increasing sequence of integers $(n_m)_{m=1}^\infty$ for which

$$|E[\nu_{\beta,n_m}] - E[\nu_{0,n_m}]| < K \text{ for all } m. \tag{1.6}$$

If $\limsup_{m\to\infty} \beta(n_m) = \beta_0 > 0$, then there exists a subsequence of $(n_m)_{m=1}^\infty$, which will be denoted the same to keep the notation simple, such that $\lim_{m\to\infty} \beta(n_m) = \beta_0$. For this subsequence, however, the difference

$$E[\nu_{\beta,n_m}] - E[\nu_{0,n_m}] \;=\; \sqrt{n_m}\Big|Q_\beta\big(Y_1^2 = (i,j)\big) - P\big(X_1^2 = (i,j)\big)\Big|$$

tends to ∞ with $m \to \infty$ because by (1.3) the absolute value converges to a positive value independent of m. This is, however, a contradiction with (1.6).

If $\lim_{m\to\infty} \beta(n_m) = 0$, one find the contradiction in a different manner. By the FI condition from Assumption 3, there must exist states (i,j) such that $\frac{d}{d\beta}Q_{\beta=0}(Y_1^2 = (i,j)) \neq 0$. From the Taylor expansion[2] of $Q_\beta(Y_1^2 = (i,j))$ at $\beta = 0$ with Lagrange remainder and $0 < \xi < 1$

$$E[\nu_{\beta,n_m}] - E[\nu_{0,n_m}] \;=\; \sqrt{n_m}\beta\left|\frac{d}{d\beta}Q_{\beta=0}\big(Y_1^2 = (i,j)\big) + \frac{1}{2}\beta\frac{d^2}{d\beta^2}Q_{\xi\beta}\big(Y_1^2 = (i,j)\big)\right|, \tag{1.7}$$

which tends to ∞ as $m \to \infty$ when $\sqrt{n_m}\beta \to \infty$, which is again a contradiction with (1.6). In summary, $E[\nu_{\beta,n}] - E[\nu_{0,n}] \to \infty$ holds for any sequence $\beta(n)$ for which $\sqrt{n}\beta(n) \to \infty$.

Lemma 1 proved below shows that exponential forgetting of Markov chains guarantees that

$$Var[\nu_{\beta,n}] < C \tag{1.8}$$

for some constant C independent of n and β. Equations (1.5) and (1.8) are all that is needed to construct detectors D_n that will put the Steganographer at risk for all sufficiently large n. The detector D_n has the following form

$$\begin{aligned} \nu_{\beta,n} > T &\quad \text{decide stego } (\beta > 0) \\ \nu_{\beta,n} \leq T &\quad \text{decide cover } (\beta = 0), \end{aligned}$$

where T is a fixed threshold. It is now shown that T can be chosen to make the detector probability of false alarms and missed detections satisfy

$$\begin{aligned} P_{FA} &< P_{FA}^* \\ P_{MD} &< P_{MD}^* \end{aligned}$$

for sufficiently large n. The threshold $T(P_{FA}^*)$ will be determined from the requirement that the probability of the right tail, $x \geq T(P_{FA}^*)$, under H_0 is at most P_{FA}^*. Using Chebyshev's inequality,

$$P_{FA} = Pr(\nu_{0,n} \geq T) \leq Pr(|\nu_{0,n}| \geq T) \;\leq\; \frac{Var[\nu_{0,n}]}{T^2} < \frac{C}{T^2}.$$

[1] For any two statements A, B, $\mathbb{I}_{\{A,B\}} = 1$ if A and B are true, otherwise $\mathbb{I}_{\{A,B\}} = 0$.

[2] The Taylor expansion is valid since by its form the function $Q_\beta(Y_k^{k+1} = (i,j))$ is analytic.

Setting $T = \sqrt{C/P^*_{FA}}$ gives $P_{FA} < P^*_{FA}$.

Because of the growing difference between the means (1.5), one can find n large enough so that the probability of the left tail, $x \le T(P^*_{FA})$, under H_1 is less than or equal to P^*_{MD}. Again, one can use the Chebyshev's inequality with the bound on the variance of $\nu_{\beta,n}$ to prove this:

$$\begin{aligned} P_{MD} &= Pr\big(\nu_{\beta,n} < T(P^*_{FA})\big) = Pr\big(\nu_{\beta,n} - E[\nu_{\beta,n} - \nu_{0,n}] < T(P^*_{FA}) - E[\nu_{\beta,n} - \nu_{0,n}]\big) \\ &\le Pr\big(|\nu_{\beta,n} - E[\nu_{\beta,n} - \nu_{0,n}]| > E[\nu_{\beta,n} - \nu_{0,n}] - T(P^*_{FA})\big) < \frac{C}{\big(E[\nu_{\beta,n} - \nu_{0,n}] - T(P^*_{FA})\big)^2}, \end{aligned}$$

which can be made arbitrarily small for sufficiently large n because $E[\nu_{\beta,n}] - E[\nu_{0,n}] \to \infty$. This establishes the first part of the square root law.

Part 2 [Asymptotic undetectability] Now it is proved that when $\sqrt{n}\beta \to 0$, then the KL divergence between the distributions of cover and stego objects tends to zero,

$$d_n(\beta) = D_{KL}\Big(P^{(n)} || Q^{(n)}_\beta\Big) = \sum_{y_1^n \in \mathcal{X}^n} P^{(n)}(X_1^n = y_1^n) \lg \frac{P^{(n)}(X_1^n = y_1^n)}{Q^{(n)}_\beta(Y_1^n = y_1^n)} \to 0, \tag{1.9}$$

which will establish that the steganography is ε-secure for any $\varepsilon > 0$ for sufficiently large n. By the well-known connection between hypothesis testing and KL divergence [9], no nontrivial upper bound on false alarms and missed detections will be met, for large enough n.

Using Taylor expansion of $d_n(\beta)$ with Lagrange remainder at $\beta = 0$, $d_n(\beta) = d_n(0) + d'_n(0)\beta + \frac{d''_n(\upsilon\beta)}{2!}\beta^2$, where $0 < \upsilon < 1$. This step is valid since under the above assumptions all derivatives of (normalized) KL divergence are continuous w.r.t. β (the complete proof appears in this technical report [18]). The term $d_n(0)$ is zero because both distributions are the same when $\beta = 0$. The term $d'_n(0)$ is also zero because

$$\begin{aligned} d'_n(0) &= \lim_{\beta\to 0} d'_n(\beta) = \lim_{\beta\to 0} \frac{-1}{\log 2} \sum_{y_1^n} P^{(n)}(X_1^n = y_1^n) \frac{\frac{d}{d\beta} Q^{(n)}_\beta(Y_1^n = y_1^n)}{Q^{(n)}_\beta(Y_1^n = y_1^n)} = \frac{-1}{\log 2} \sum_{y_1^n} \frac{d}{d\beta} Q^{(n)}_{\beta=0}(Y_1^n = y_1^n) \\ &= \lim_{\beta\to 0} \frac{-1}{\log 2} \frac{d}{d\beta} \Big(\underbrace{\sum_{y_1^n} Q^{(n)}_\beta(Y_1^n = y_1^n)}_{=1} \Big) = 0. \end{aligned}$$

Finally, by Lemma 2 in Appendix there exists a constant $\tilde{C}$, such that $\frac{1}{n} d''_n(\beta) < \tilde{C}$ for $\beta \in [0, \beta_0]$ and all n. Thus, $d_n(\beta) \le \frac{1}{2}\tilde{C} n \beta^2 \to 0$ when $\sqrt{n}\beta \to 0$.

Part 3 [Asymptotic $C\epsilon^2$-security] To prove the third part of the square root law, one again expands the KL divergence $d_n(\beta)$ at $\beta = 0$ up to the third order with the Lagrange form of the remainder

$$d_n(\beta) = \frac{1}{2!}\Big(\frac{d''_n(0)}{n}\Big) n\beta^2 + \frac{1}{3!}\Big(\frac{d'''_n(\upsilon\beta)}{n}\Big) n\beta^3 \tag{1.10}$$

for some $0 < \upsilon < 1$. According to [18], both normalized derivatives of the KL divergence, $\frac{1}{n} d''_n(0)$ and $\frac{1}{n} d'''_n(\upsilon\beta)$, are upper bounded by the same finite constant $\tilde{C}$ for all $\beta \in [0, \beta_0]$. Since $\beta(n)\sqrt{n} \to \epsilon$ with $n \to \infty$, $\beta(n) \to 0$ and thus the expansion is valid. By the same reason, the second term in (1.10) converges to zero as $n \to \infty$. From this result, one obtains the asymptotic bound on KL divergence in the form $d_n(\beta) \le \frac{1}{2}\tilde{C}\epsilon^2$ as was to be shown.

1.3. **Discussion.** A general theme is now emerging in steganography literature: whether steganography is performed in a large batch of cover objects or a single large object, there is a wide range of situations in which *secure capacity grows according to the square root of the cover size.* Such results will likely hold for all stegosystems that are not perfectly secure in the sense of Cachin. It appears that the theory of *hidden* information is quite unlike the traditional theory of information.

The result presented here is the first to allow dependence between the components of the cover. The square root law of steganographic capacity was proved for single covers under essentially three conditions: that they can be represented as a Markov chain, that the embedding operation can be modeled as independent substitutions of one state for another, and that the embedding scheme does not preserve all statistical properties of the cover. This applies to a very wide range of popular steganographic algorithms, in spatial and transform domains. The last condition is important because it is known that perfectly secure steganography, conveying information at a linear rate, can always be constructed if the cover source is perfectly understood [89]. However, it can be argued that digital media cover sources will never be perfectly understood. To explain why it makes sense to assume that the warden has perfect knowledge of the covers (necessary for construction

of the detectors in Part 1 of the proof), one needs to look at the cautious nature of Kerckhoffs' principle: although the steganographer may believe that the warden does not know the complete cover distribution, there is always the risk that the warden knows more than the steganographer does. For example, she might know more about dependencies between cover elements. Since the steganographer cannot say for certain how much the warden will know, they must assume the worse case: complete knowledge.

The formulation of the theorem parallels that of [51]: embedding at a rate faster than $\sqrt{n}$ leads to eventual detection, whereas embedding at a rate slower than $\sqrt{n}$ leads to eventual ε-security. At rates $A\sqrt{n}$, the stegosystem is ϵ-secure. This does *not* refer to embedding at a diminishing rate in a single cover object (which would be a different theorem, and is an avenue for future research). The quantity $\beta(n)$ describes a constant embedding rate throughout an object of size n: one could think of the function β as giving a strategy describing how much data can be hidden in an object of each size. The SRL says that over-ambitious strategies lead to easier detection in larger objects, cautious strategies lead to more difficult detection in larger objects, and quantifies the boundary.

The square root law has some important implications in steganography and steganalysis. Most significantly, the simple fact that capacity is sublinear means that a true steganographic channel, with positive rate, cannot be constructed unless the cover source is known perfectly. For steganalysis, the SRL explains in part why the same relative payload can be detected more accurately in large images (however it is not the whole explanation: larger images tend to have more smooth gradients and less noise, and these factors also influence detection accuracy). Thus, when benchmarking steganography, the distribution of image sizes in the database influences the reliability of steganalysis and makes it more difficult to compare the results on two different databases. To resolve this issue, one might switch to measuring the payload in bits per square root of pixel, an idea further explained in Section 2.3.

Finally, it should be emphasized that the square root law of capacity relates to the number of changes caused by the embedding process, and not to the size of the information transmitted. With adaptive source coding methods (even simple matrix embedding based on Hamming codes will do) the number of bits of information which can be conveyed, by making at most c changes in n cover locations, is $O(c\log(n/c))$. Therefore the SRL implies an asymptotic information capacity which is $O(\sqrt{n}\log n)$ (i.e., still sublinear), in the absence of perfect steganography.

1.4. **Proofs.** In this appendix, the PI includes two auxiliary lemmas needed in the proof of the SRL in Section 1.2.

Lemma 1: Let $\nu_{\beta,n}$ be the random variable defined in (1.4) for a fixed value of the parameter β and cover size n. The variance of this random variable can be bounded by a constant C for every value of β and n

$$\exists C, \forall\beta, \forall n \quad Var[\nu_{\beta,n}] \leq C.$$

Proof: From the definition of $\nu_{\beta,n}$

$$\begin{aligned}\frac{(n-1)^2}{n}Var[\nu_{\beta,n}] &= E\Big[\Big(\sum_{k=1}^{n-1}\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}\Big)^2\Big] - E\Big[\sum_{k=1}^{n-1}\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}\Big]^2 \leq \sum_{k=1}^{n-1}Var[\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}] + \\ &\quad +2\Big[\sum_{k+1<\hat{k}}E[\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}] - E[\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}]E[\mathbb{I}_{\{Y_k^{k+1}=(i,j)\}}]\Big] + 2n. \qquad (1.11)\end{aligned}$$

In the last sum, the PI bounded all terms for $k = \hat{k} - 1$ by 1 and thus obtained the term $2n$ in the last inequality. For any event A,

$$Var[\mathbb{I}_A] = Pr(A) - Pr(A)^2 = \tfrac{1}{4} - \big(Pr(A) - \tfrac{1}{2}\big)^2 \leq \tfrac{1}{4}$$

so the first term is bounded by $\frac{n-1}{4}$.

Finally, one finds an upper bound on the sum in (1.11) in the form of $C_2 n$ for some positive constant C_2. This will give us the proof because $Var[\nu_{\beta,n}] \leq \frac{n}{(n-1)^2}((n-1)\frac{1}{4} + 2C_2 n + 2n) \leq 4(\frac{1}{4} + 2C_2 + 2)$, and $\frac{n^2}{(n-1)^2} \leq 4$ for $n \geq 2$. Thus $C = 8C_2 + 9$.

The PI starts by showing that

$$\begin{aligned}&Q_\beta(Y_k^{k+1} = (i,j), Y_{\hat{k}}^{\hat{k}+1} = (i,j)) - Q_\beta(Y_k^{k+1} = (i,j))Q_\beta(Y_{\hat{k}}^{\hat{k}+1} = (i,j)) \\ &= \Big[\underbrace{Q_\beta(Y_{\hat{k}}^{\hat{k}+1} = (i,j)|Y_k^{k+1} = (i,j)) - Q_\beta(Y_{\hat{k}}^{\hat{k}+1} = (i,j))}_{\leq N^2\rho^{\hat{k}-k-2}}\Big]Q_\beta(Y_k^{k+1} = (i,j)) \leq N^2\rho^{\hat{k}-k-2}, \qquad (1.12)\end{aligned}$$

for some $0 \leq \rho < 1$ and $k+1 < \hat{k}$ (N is the number of all possible states of the MC). In other words, the HMC is exponentially forgetting its initial condition. Then, one will be able to bound the sum in (1.11) by $N^2 \sum_{\hat{k}=3}^{n} \sum_{k=1}^{\hat{k}-2} \rho^{\hat{k}-k-2} = N^2 \sum_{\hat{k}=3}^{n} \frac{1-\rho^{\hat{k}-2}}{1-\rho} \leq N^2 \sum_{\hat{k}=3}^{n} \frac{1}{1-\rho} = N^2(n-2)\frac{1}{1-\rho} \leq \frac{N^2 n}{1-\rho}$. Thus, $C_2 = \frac{N^2}{1-\rho}$ because $Q_\beta\big(Y_k^{k+1} = (i,j)\big) \leq 1$.

The term $Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j)\big)$ in (1.12) can be written as

$$Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j)\big) = \sum_{(\hat{i},\hat{j})} Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j) \big| X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j})\big) P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j})\big) = \sum_{(\hat{i},\hat{j})} b_{\hat{i},i} b_{\hat{j},j} P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j})\big). \tag{1.13}$$

The term $Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j) \big| Y_k^{k+1} = (i,j)\big)$ in (1.12) can be written as

$$\begin{aligned} Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j) \big| Y_k^{k+1} = (i,j)\big) &= \frac{Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j), Y_k^{k+1} = (i,j)\big)}{Q_\beta\big(Y_k^{k+1} = (i,j)\big)} \\ &= \frac{\sum_{(\hat{i},\hat{j})} \sum_{(\tilde{i},\tilde{j})} b_{\hat{i},i} b_{\hat{j},j} b_{\tilde{i},i} b_{\tilde{j},j} P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}), X_k^{k+1} = (\tilde{i},\tilde{j})\big)}{Q_\beta\big(Y_k^{k+1} = (i,j)\big)} = (\#). \end{aligned}$$

Finally, $P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}), X_k^{k+1} = (\tilde{i},\tilde{j})\big)$ can be factorized as $P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_k^{k+1} = (\tilde{i},\tilde{j})\big) P\big(X_k^{k+1} = (\tilde{i},\tilde{j})\big)$. Due to the Markov property of the random variable X_1^n, $P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_k^{k+1} = (\tilde{i},\tilde{j})\big) = P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_{k+1} = \tilde{j}\big)$. For each pair of indices $(\hat{i},\hat{j})$, the PI defines the index $\tilde{j}_{max} = \arg\max_{\tilde{j}} P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_{k+1} = \tilde{j}\big)$. Then

$$\begin{aligned} (\#) &\leq \frac{\sum_{(\hat{i},\hat{j})} b_{\hat{i},i} b_{\hat{j},j} P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_{k+1} = \tilde{j}_{max}\big) \sum_{(\tilde{i},\tilde{j})} b_{\tilde{i},i} b_{\tilde{j},j} P\big(X_k^{k+1} = (\tilde{i},\tilde{j})\big)}{Q_\beta\big(Y_k^{k+1} = (i,j)\big)} \\ &\overset{(1.13)}{=} \sum_{(\hat{i},\hat{j})} b_{\hat{i},i} b_{\hat{j},j} P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_{k+1} = \tilde{j}_{max}\big). \end{aligned} \tag{1.14}$$

Now, one can combine (1.13) and (1.14) to prove (1.12) as

$$\begin{aligned} &Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j) \big| Y_k^{k+1} = (i,j)\big) - Q_\beta\big(Y_{\hat{k}}^{\hat{k}+1} = (i,j)\big) \\ &\overset{(1.13),(1.14)}{\leq} \sum_{(\hat{i},\hat{j})} b_{\hat{i},i} b_{\hat{j},j} \Big(P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j}) | X_{k+1} = \tilde{j}_{max}\big) - P\big(X_{\hat{k}}^{\hat{k}+1} = (\hat{i},\hat{j})\big) \Big) \\ &= \sum_{(\hat{i},\hat{j})} b_{\hat{i},i} b_{\hat{j},j} P\big(X_{\hat{k}+1} = \hat{j} | X_{\hat{k}} = \hat{i}\big) \Big(P\big(X_{\hat{k}} = \hat{i} | X_{k+1} = \tilde{j}_{max}\big) - P\big(X_{\hat{k}} = \hat{i}\big) \Big) \leq N^2 \rho^{\hat{k}-k-2}. \end{aligned} \tag{1.15}$$

It is a well known result in MCs that the absolute value of the term $P\big(X_{\hat{k}} = \hat{i} | X_{k+1} = \tilde{j}_{max}\big) - P\big(X_{\hat{k}} = \hat{i}\big)$ in (1.15) can be bounded by $\rho^{\hat{k}-k-2}$ (exponential forgetting), for some constant $0 \leq \rho < 1$. This is because the MC is irreducible due to the assumption $a_{i,j} \geq \delta$ (see equation (2.2) on page 173 in Doob [17]). This bound does not depend on $\tilde{j}_{max}$. The final bound does not depend on β because $b_{\hat{i},i} \leq 1$ and $b_{\hat{j},j} \leq 1$.

Lemma 2: Let $d_n(\beta) = D_{KL}(P^{(n)} || Q_\beta^{(n)})$ be the KL divergence between n-element cover and stego sources as defined in (1.9). Then,

$$\exists \beta_0, \exists C > 0, \forall \beta \in [0, \beta_0], \forall n \quad \frac{1}{n} d_n''(\beta) < \tilde{C}.$$

In other words, the second derivative of the normalized KL divergence, $\frac{1}{n} d_n''(\beta)$, can be bounded by a constant $\tilde{C}$ for each n and β. And this bound does not depend on n or β.

Proof: The problem of bounding normalized derivatives of KL divergence for the case of HMC was studied by Mevel et al. [60]. Their results, namely Theorem 4.4 and Theorem 4.7, however, cannot be directly applied to our case because our assumptions are different. In particular, Assumption C on page 1124 is not satisfied because zeros are allowed in matrix $\mathbb{B}$. Motivated by this work, the PI derives a more general result about the normalized KL divergence and its derivatives (see the report in [18]. Intuitively, one can expect the normalized KL divergence to be arbitrarily smooth and bounded due to the smooth transition from P to Q_β and the fact that $d_n(0) = 0$. The main result of the report, formally stated in [18], Theorem 3, says that every derivative of $\frac{1}{n} d_n(\beta)$ w.r.t. β (and the function $\frac{1}{n} d_n(\beta)$ itself) is uniformly bounded

and Lipschitz-continuous (or simply continuous) on $[0, \beta_0]$. These properties are independent of $n \geq 1$. From this fact, Lemma 2 can be obtained as a special case. This result also allows us to expand the KL divergence into a Taylor series with respect to β.

2. Complete characterization of perfectly secure stego-systems with mutually independent embedding operation

In steganography, the sender and receiver communicate by hiding their messages in generally trusted media, such as digital images, so that one cannot distinguish between the original (cover) objects and the objects carrying the message (stego objects). Formally, the security of a stego-system is evaluated using the Kullback-Leibler divergence between the distributions of cover and stego objects [9]. Systems with zero KL divergence are called perfectly secure.

Formally, a stego-system is a combination of an embedding algorithm and a cover source. The vast majority of practical stego-systems hide messages by modifying individual cover elements using mutually independent embedding operations, e.g., LSB and ±1 embedding, F5 algorithm, perturbed quantization, MMx, stochastic modulation, and many others (see [37] and the references therein).

This section of the report provides a complete characterization of perfectly secure stego-systems for the class of embedding algorithms that employ mutually independent (MI) embedding operations, which is the class of operations for which the SRL was proved in the previous section. The cover distributions of all perfectly secure systems form a linear vector space spanned by distributions determined by the embedding operation. Moreover, perfect security (zero KL divergence) is equivalent to satisfying a simple condition related to Fisher information. This result suggests that Fisher information can be used as an *equivalent* descriptor of steganographic security.

Definition 1. Steganography is *perfectly secure* iff

$$d(\beta) \triangleq D_{KL}(P||Q_\beta) = \sum_{y_1^n \in \mathcal{X}^n} P(y_1^n) \log \frac{P(y_1^n)}{Q_\beta(y_1^n)} = 0,$$

or *ϵ-secure* if $d(\beta) \leq \epsilon$.

For better flow, the PI reminds that the impact of embedding with parameter $\beta \in [0, \beta_0]$ on the k-th element can be captured using the matrix $b_{i,j}(\beta) \triangleq Pr(Y_k = j | X_k = i) = \delta_{i,j} + \beta c_{i,j}$, for some constants $c_{i,j} \geq 0$ for $i \neq j$, $c_{i,i} = -\sum_j c_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. In a matrix form, $\mathbb{B}_\beta = \mathbb{I} + \beta\mathbb{C}$, where $\mathbb{B}_\beta \triangleq (b_{i,j}(\beta))$, $\mathbb{I}$ is the identity matrix, and $\mathbb{C} \triangleq (c_{i,j})$. It is also assumed that embedding operations are mutually independent, $Pr(Y_1^n | X_1^n) = \prod_{k=1}^n Pr(Y_k | X_k)$. By the definition of $b_{i,j}$, the matrix $\mathbb{B}_\beta$ is stochastic, $\sum_j b_{i,j} = 1$. Finally, it is assumed that $b_{i,i}(\beta) > 0$ for all $\beta \in [0, \beta_0]$. The matrix $\mathbb{B}_\beta$ represents an embedding algorithm with MI embedding operation (simply MI embedding). Examples of practical embedding methods that fall under this framework are shown in Figure 2.1.

To simplify the language in this report, one will speak of security of a cover source w.r.t. a given MI embedding meaning that the *cover source is perfectly secure w.r.t.* $\mathbb{B}$, if the resulting stego-system is perfectly secure. It does then make sense to inquire about all possible perfectly secure cover sources w.r.t. MI embedding with matrix $\mathbb{B}_\beta$.

2.1. A few known results from ergodic theory.

Some results from the theory of ergodic classes are now reviewed [17]. They will be later applied to the stochastic matrix $\mathbb{B}_\beta$. For states $i, j \in \mathcal{X}$, j is called *a consequent* of i (of order k) ($i \to j$) iff $\exists k, (\mathbb{B}_\beta^k)_{i,j} \neq 0$. State $i \in \mathcal{X}$ is *transient* if it has a consequent of which it is not itself a consequent, i.e., $\exists j \in \mathcal{X}$ such that $(i \to j) \Rightarrow (j \not\to i)$. Furthermore, $i \in \mathcal{X}$ is *non-transient* if it is a consequent of every one of its consequents, $\forall j \in \mathcal{X}$, $(i \to j) \Rightarrow (j \to i)$. The set $\mathcal{X}$ can be decomposed as $\mathcal{X} = \mathcal{F} \cup \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k$, where $\mathcal{F}$ is the set of all transient states and $\mathcal{E}_a$, $a \in \{1, \ldots, k\}$, are so called ergodic classes. Two non-transient states are put into one ergodic class if they are consequents of each other.

Let matrix $\mathbb{B}_\beta$ have k ergodic classes. Then, there exist k linearly independent left eigenvectors, denoted as $\pi^{(1)}, \ldots, \pi^{(k)}$, of matrix $\mathbb{B}_\beta$ corresponding to eigenvalue 1, called *invariant distributions*. If $\pi^{(a)}\mathbb{B}_\beta = \pi^{(a)}$, for some $a \in \{1, \ldots, k\}$, then $\pi_i^{(a)} > 0$ for all $i \in \mathcal{E}_a$, and $\pi_i^{(a)} = 0$ otherwise. Every other π satisfying $\pi\mathbb{B}_\beta = \pi$ is obtained by a convex linear combination of $\{\pi^{(a)} | a \in \{1, \ldots, k\}\}$. For a complete reference, see [17, Chapter V, §2]. The set of ergodic classes for matrix $\mathbb{B}_\beta$ depends only on the set $\{(i,j) | b_{i,j}(\beta) \neq 0\}$. Since $b_{i,j}(\beta) = 0$ iff $c_{i,j} = 0$ for $i \neq j$ and $b_{i,i}(\beta) > 0$ for $\beta \in (0, \beta_0]$, the structure of ergodic classes does not depend on β. Moreover, if $\pi\mathbb{B}_\beta = \pi$ for some $\beta > 0$, then $\pi\mathbb{C} = 0$ and thus all invariant distributions are independent of β, because $\pi\mathbb{B}_{\beta'} = \pi\mathbb{I} + \beta'\pi\mathbb{C} = \pi\mathbb{I} = \pi$. By this reason, the index β will be frequently omitted.

2.2. Perfectly secure cover sources under mutually independent embedding operation.

The matrix $\mathbb{B}$ represents an arbitrary MI embedding with k ergodic classes $\mathcal{E}_a$ and invariant distributions $\pi^{(a)}$, $a \in \{1, \ldots, k\}$. The following example describes a construction of perfectly secure cover sources w.r.t. $\mathbb{B}$.

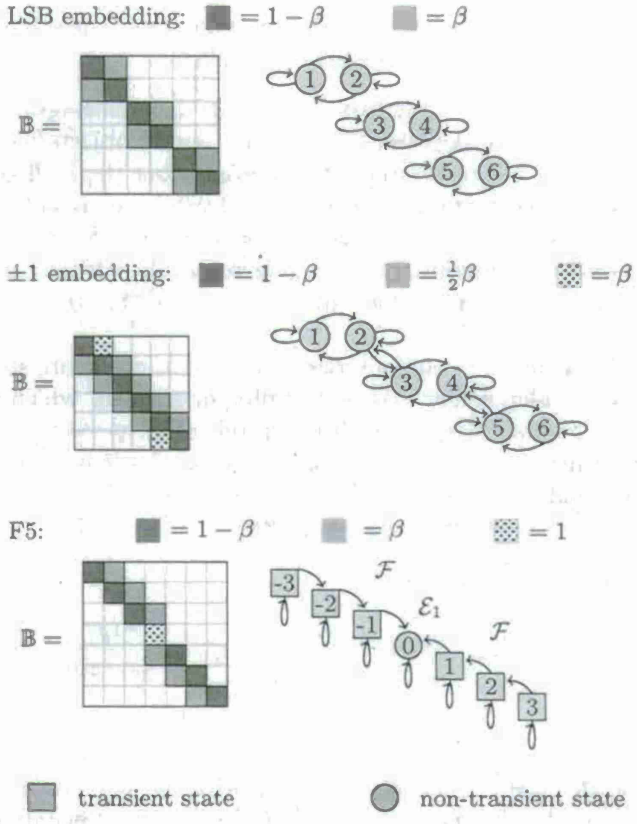

FIGURE 2.1. Examples of several embedding methods and their ergodic classes.

Example 2. *[Perfectly secure cover sources]* Let $P^{(2)}$ be a probability distribution on 2-element cover objects defined as $P^{(2)}(X_1^2 = (i,j)) = \pi_i^{(a)}\pi_j^{(b)}$ for some $a, b \in \{1, \ldots, k\}$. Then $P^{(2)}$ is a perfectly secure cover source w.r.t. $\mathbb{B}$ because

$$\begin{aligned} \&Q\beta(2)\big(Y12 = (i,j)\big) &= \Big(\sum \hat{\imath} b\hat{\imath}, iP(X1 = \hat{\imath})\Big)\Big(\sum \hat{\jmath} b\hat{\jmath}, jP(X2 = \hat{\jmath})\Big) \\ \& &= \big(\pi(a)\mathbb{B}\big)i\big(\pi(b)\mathbb{B}\big)j = \pi i(a)\pi j(b) = P(2)\big(X12 = (i,j)\big), \end{aligned}$$

and thus both distributions $P^{(2)}$, and $Q_\beta^{(2)}$ are identical, which implies perfect security. Since this construction does not depend on the particular choice of $a, b \in \{1, \ldots, k\}$, one can create k^2 perfectly secure cover sources w.r.t. $\mathbb{B}$. The probability distributions $P^{(2)}$ obtained from this construction are linearly independent and form a k^2-dimensional linear vector space. By a similar construction, one can construct k^n n-element linearly independent perfectly secure cover sources w.r.t. $\mathbb{B}$.

It is shown next that there are no other linearly independent perfectly secure cover sources w.r.t. $\mathbb{B}$.

Theorem 3. [Mutually independent embedding] *There are exactly k^n linearly independent perfectly secure probability distributions P on n-element covers. Every perfectly secure probability distribution P w.r.t. $\mathbb{B}$ can be obtained by a convex linear combination of k^n linearly independent perfectly secure distributions described in Example 2.*

Proof. It is sufficient to prove that there cannot be more than k^n linearly independent perfectly secure probability distributions P on n-element covers. The proof is shown for $n = 2$ and then its generalization is discussed.

Define the following matrices $\mathbb{P} \triangleq (p_{i,j})$, $p_{i,j} = P(X_1^2 = (i,j))$, and $\mathbb{Q} \triangleq (q_{i,j})$, $q_{i,j} = Q_\beta(Y_1^2 = (i,j))$. By defininition of MI embedding,

$$\begin{aligned} q_{ij} &= \sum_{(v,w)\in\mathcal{X}^2} Q_\beta(Y_1^2 = (i,j)|X_1^2 = (v,w))P(X_1^2 = (v,w)) \\ &= \sum_{v,w\in\mathcal{X}} b_{vi}b_{wj}p_{vw}. \end{aligned}$$

Define matrix $\mathbb{D} \triangleq (d_{u_1^2,v_1^2})$ of size $N^2 \times N^2$, where $d_{u_1^2,v_1^2} = b_{u_1,v_1} b_{u_2,v_2}$. If $\vec{p}$ is defined as one big row vector of elements $p_{i,j}$ and similarly $\vec{q}$, then assuming perfect security of cover source w.r.t. $\mathbb{B}$ ($\mathbb{P} = \mathbb{Q}$), one obtains $\vec{q} = \vec{p}\mathbb{D} = \vec{p}$ and thus $\vec{p}$ is a left eigenvector of $\mathbb{D}$ corresponding to 1. Matrix $\mathbb{D}$ is stochastic and thus it is sufficient to show that it has k^2 ergodic classes.

This is achieved by showing first that

$$u_1^2 \overset{(m)}{\to} v_1^2 \Leftrightarrow (u_1 \overset{(m)}{\to} v_1) \text{ and } (u_2 \overset{(m)}{\to} v_2),\ u_1^2, v_1^2 \in \mathcal{X}^2. \tag{2.1}$$

By $u_1^2 \overset{(m)}{\to} v_1^2$ it is meant that v_1^2 is a consequent of u_1^2 of order m in terms of matrix $\mathbb{D}$. If $u_1^2 \overset{(m)}{\to} v_1^2$, then there exist $m-1$ intermediate states ${}_1w_1^2, \ldots, {}_{m-1}w_1^2$, such that $d_{u,{}_1w} d_{{}_1w,{}_2w} \cdots d_{{}_{m-1}w,v} > 0$. Since $d_{u_1^2,v_1^2} = b_{u_1,v_1} b_{u_2,v_2}$, this implies the existence of both paths $u_i \overset{(m)}{\to} v_i$ of order m, $i = 1, 2$. The converse is true by the same reason.

It is now shown that $\mathcal{E}_a \times \mathcal{E}_b$, $a, b \in \{1, \ldots, k\}$ are the only ergodic classes. If $u_1 \overset{(m_1)}{\to} v_1$ and $u_2 \overset{(m_2)}{\to} v_2$, then $u_1^2 \overset{(m_1+m_2)}{\to} v_1^2$ for all $u_1, v_1 \in \mathcal{E}_a$ and $u_2, v_2 \in \mathcal{E}_b$, because the path from u_i to v_i can be arbitrarily extended by adding self loops of type $j \to j$ since all diagonal terms $b_{j,j}$ are positive and thus by (2.1) one obtains $u_1^2 \overset{(m_1+m_2)}{\to} v_1^2$. Finally by $u_1, v_1 \in \mathcal{E}_a$ and $u_2, v_2 \in \mathcal{E}_b$, $v_i \to u_i$ and by the same argument $v_1^2 \to u_1^2$, and therefore $\mathcal{E}_a \times \mathcal{E}_b$ are ergodic classes. Any other state $u_1^2 \in \mathcal{E}_a \times \mathcal{F} \cup \mathcal{F} \times \mathcal{E}_a \cup \mathcal{F} \times \mathcal{F}$ must be transient w.r.t. $\mathbb{D}$, otherwise by (2.1) a contradiction with $u_i \in \mathcal{F}$ for some i is obtained.

This proof can be generalized for $n \geq 3$ by proper definition of matrices $\mathbb{P}$, $\mathbb{Q}$, and $\mathbb{D}$. In general, matrix $\mathbb{D}$ has size $N^n \times N^n$. By similar construction one obtains k^n ergodic classes of generalized matrix $\mathbb{D}$. However, k^n linearly independent distributions are already known. □

2.3. **Perfect security and Fisher information.** Here, it is shown that for stego-systems with MI embedding perfect security can be captured using Fisher information. From Taylor expansion of KL divergence, for small β, $d(\beta) = \frac{1}{2}\beta^2 I(0) + O(\beta^3)$ where $I(0) = \partial^2 d(\beta)/\partial\beta^2|_{\beta=0}$ is the Fisher information w.r.t. β. If for some stego-system $d(\beta) = 0$ for $\beta \in [0, \beta_0]$, then $I(0) = 0$ from the Taylor expansion. Even though the opposite does not hold in general, the PI will prove that for MI embedding zero Fisher information implies perfect security. In other words, a stego-system with MI embedding is perfectly secure for $\beta \in [0, \beta_0]$ if and only if $I(0) = 0$. This supplies a simpler condition for verifying perfect security than the KL divergence. Fisher information also provides a connection to quantitative steganalysis because $1/I(\beta)$ is the lower bound on variance of unbiased estimators of β. Moreover, $I(0)$ could be used for comparing (benchmarking) stego-systems.

First, he condition $I(0) = 0$ is reformulated:

Proposition 4. *Let P and Q_β be probability distributions of cover and stego objects with n elements embedded with parameter β. The Fisher information is zero if and only if the FI-condition is satisfied*

$$\forall y_1^n \in \mathcal{X}^n \quad \Big(P(X_1^n = y_1^n) > 0\Big) \Rightarrow \Big(\frac{d}{d\beta} Q_\beta(y_1^n)\big|_{\beta=0} = 0\Big). \tag{2.2}$$

Proof. The second derivative of $d(\beta)$ at β, $d''(\beta)$, can be written as

$$I(\beta) = -\sum_{y_1^n \in \mathcal{X}^n} P(y_1^n)\Big(\frac{Q''_\beta(y_1^n)}{Q_\beta(y_1^n)} - \Big(\frac{Q'_\beta(y_1^n)}{Q_\beta(y_1^n)}\Big)^2\Big), \tag{2.3}$$

where $Q'_\beta(y_1^n) = \frac{\partial}{\partial\beta} Q_\beta(y_1^n)$. By $P(y_1^n) = Q_{\beta=0}(y_1^n)$, the first term in the bracket in (2.3) sums to zero at $\beta = 0$, and thus $I(0)$ is zero iff $Q'_\beta(y_1^n)\big|_{\beta=0} = 0$ is zero for all $y_1^n \in \mathcal{X}^n$ for which $P^{(n)}(y_1^n) > 0$ as was to be proved. Here, it is assumed that the KL divergence $d(\beta)$ be continuous w.r.t. β, which is valid by the construction of the matrix $\mathbb{B}$. □

The next theorem shows that the FI condition (2.2) is equivalent with perfect security for MI embedding.

Theorem 5. [Fisher information condition] *There are exactly k^n linearly independent probability distributions P on n-element covers satisfying the FI condition* (2.2). *These distributions are perfectly secure w.r.t. $\mathbb{B}$. Every other probability distribution P satisfying* (2.2) *can be obtained by a convex linear combination of k^n linearly independent perfectly secure distributions.*

Proof. From Example 2, k^n linearly independent perfectly secure distributions are available. By Taylor expansion of $d(\beta)$, these distributions satisfy the FI condition, because $d(\beta) = 0 \Rightarrow I(0) = 0$. It is sufficient to show that there cannot be more linearly independent distributions satisfying the FI condition.

Similarly as in the previous proof, the theorem is reformulated as an eigenvector problem so that one can use ergodic class theory to give the exact number of left eigenvectors corresponding to 1. Again, the proof is first carried out for the case $n = 2$.

If P satisfies (2.2), then the linear term in the Taylor expansion of $Q_\beta(y_1^2)$ w.r.t. β is zero. By the independence property, $(Q(y_1^n|x_1^n) = \prod_{i=1}^n Q(y_i|x_i))$, and the form of matrix $\mathbb{B}$ ($\mathbb{B}_\beta = \mathbb{I} + \beta\mathbb{C}$), condition (2.2) has the following form

$$\begin{aligned} \frac{dQ_\beta(y_1^2)}{d\beta}\Big|_{\beta=0} &= \lim_{\beta\to 0} \sum_{x_1^2\in\mathcal{X}^2} P(x_1^2)\frac{d}{d\beta}\prod_{i=1}^2 Q_\beta(y_i|x_i) \\ &= \sum_{x_1\in\mathcal{X}} c_{x_1,y_1}P(x_1,y_2) + \sum_{x_2\in\mathcal{X}} c_{x_2,y_2}P(y_1,x_2) = 0. \end{aligned} \tag{2.4}$$

Define matrix $\mathbb{P} \triangleq (p_{i,j})$ as $p_{i,j} = P(X_1^2 = (i,j))$ and represent it as a row vector $\vec{p}$. If one defines matrix $\mathbb{D} \triangleq (q_{u_1^2,v_1^2})$ of size $N^2 \times N^2$ as

$$d_{u_1^2,v_1^2} = \begin{cases} c_{u_1,v_1} & \text{if } u_1 \neq v_1 \text{ and } u_2 = v_2 \\ c_{u_2,v_2} & \text{if } u_1 = v_1 \text{ and } u_2 \neq v_2 \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

and a diagonal matrix $\mathbb{G} \triangleq (g_{u_1^2,v_1^2})$ of size $N^2 \times N^2$ as $g_{u_1^2,u_1^2} = -c_{u_1,u_1} - c_{u_2,u_2}$, then equation (2.4) can be written in a compact form as $\vec{p}\mathbb{D} = \vec{p}\mathbb{G}$. Both matrices $\mathbb{D}$ and $\mathbb{G}$ are non-negative by their definitions. Let $\mathbb{H} = \mathbb{I} + \gamma(\mathbb{D} - \mathbb{G})$. Using $\gamma = (\max_{u_1^2\in\mathcal{X}^2} g_{u_1^2,u_1^2})^{-1}$, the matrix $\mathbb{H}$ is stochastic and $\vec{p}\mathbb{H} = \vec{p}$ iff $\vec{p}\mathbb{D} = \vec{p}\mathbb{G}$. Thus, (2.2) is equivalent with an eigenvalue problem for matrix $\mathbb{H}$.

First, observe that for $i \neq j$ $c_{ij} > 0$ iff $h_{(i,a),(j,a)} > 0$ for all $a \in \mathcal{X}$, because by (2.5) $h_{(i,a),(j,a)} = \gamma d_{(i,a),(j,a)} = \gamma c_{ij}$ (the first case when $u_2 = v_2$). Similarly, for $i \neq j$ $c_{ij} > 0$ iff $h_{(a,i),(a,j)} > 0$ for all $a \in \mathcal{X}$ (the second case when $u_1 = v_1$). This means that $i \to j$ iff $(i,a) \to (j,a)$ w.r.t. $\mathbb{H}$ for all $a \in \mathcal{X}$ and similarly $i \to j$ iff $(a,i) \to (a,j)$ w.r.t. $\mathbb{H}$ for all $a \in \mathcal{X}$. This can be proved by using the previous statement. By this rule used for a given $u_1^2 \in \mathcal{E}_a \times \mathcal{E}_b$, one obtains $u_1^2 \to v_1^2$ and $v_1^2 \to u_1^2$ for all $v_1^2 \in \mathcal{E}_a \times \mathcal{E}_b$ and thus $\mathcal{E}_a \times \mathcal{E}_b$ is an ergodic class w.r.t. $\mathbb{H}$. Te PI now shows that there can not be more ergodic classes and thus all k^2 of them are found. If $u_1^2 \in \mathcal{F} \times \mathcal{E}$, then u_1^2 has to be transient w.r.t. $\mathbb{H}$, otherwise one would obtain contradiction with $u_1 \in \mathcal{F}$. This is because the only consequents of order 1 are of type $(i,a) \to (j,a)$ or $(a,i) \to (a,j)$, therefore if $u_1^2 \in \mathcal{F} \times \mathcal{E}$, one chooses $v_1^2 \in \mathcal{X} \times \mathcal{E}$, such that $v_1 \not\to u_1$ (u_1 is transient and thus such v_1 must exist). State u_1^2 must be transient otherwise $u_1^2 \leftrightarrow v_1^2$ implies $u_1 \leftrightarrow v_1$, which results in contradiction with $v_1 \not\to u_1$. Similarly for $u_1^2 \in \mathcal{E} \times \mathcal{F} \cup \mathcal{F} \times \mathcal{F}$. This proof can be generalized for $n \geq 3$ by assuming larger matrices $\mathbb{P}$, $\mathbb{D}$, $\mathbb{G}$, and $\mathbb{H}$, obtaining exactly k^n linearly independent perfectly secure distributions satisfying the FI condition. □

Next, the PI discusses the structure of the set of invariant distributions for a given MI embedding and shows how to find ergodic classes from matrix $\mathbb{B}$ in practice. By Theorem 2.1 from [17, Chapter V, page 175], this can be done by inspecting the matrix limit $\mathbb{M} = (m_{i,j}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{B}^i$. According to this theorem, state i is non-transient iff $m_{i,i} > 0$ and is transient otherwise. Two non-transient states $i, j \in \mathcal{X}$ are included in one ergodic class if $m_{i,j} > 0$. All rows of the matrix $\mathbb{M}$ corresponding to states in one ergodic class $\mathcal{E}_a$ are the same and equal to the invariant distribution of this class, $\pi^{(a)}$.

This section is closed with a short discussion of two practical embedding algorithms. For the F5 embedding algorithm [90], the set of states $\mathcal{X} = \{-1024, \ldots, 1024\}$. By the nature of the embedding changes (flip towards 0), there is only one ergodic set $\mathcal{E}_1 = \{0\}$ and $\mathcal{F} = \mathcal{X} \setminus \{0\}$. Thus, there is only one invariant distribution, $\pi_0 = 1$ and zero otherwise. Obviously, no message can be embedded in covers with this singular distribution.

For the case of LSB embedding over $\mathcal{X} = \{0, \ldots, 255\}$, one has $\mathcal{E}_a = \{2a, 2a+1\}$ for $a \in \{0, \ldots, 127\}$, $\mathcal{F} = \emptyset$ and $\pi_{2a}^{(a)} = \pi_{2a+1}^{(a)} = \frac{1}{2}$ and zero otherwise (LSB embedding cannot be detected in images with evened out histogram bins). Thus, sources realized as a sequence of mutually independent random variables with such a distribution are the only perfectly secure sources w.r.t. LSB embedding. Figure 2.1 shows examples of matrices $\mathbb{B}$ and ergodic classes of several known algorithms with MI embedding operation.

2.4. **Application to Markov cover sources.** In this section, the results obtained so far are reformulated for a special type of cover sources that can be modeled as first-order stationary Markov Chains (MC). The results play a key role in proving the square root law of steganographic capacity of imperfect stego-systems for Markov covers [27, 54].

First, for stationary cover sources Theorem 3 leads to this immediate corollary.

Corollary 6. *There are exactly k (instead of k^n) linearly indpendent perfectly secure stationary cover sources. These sources are i.i.d. with some invariant distribution $\pi_a, a \in 1, \ldots, k$.*

The next corollary states that in order to study perfect security of n-element stationary MC covers, it is enough to study only 2-element covers.

Corollary 7. *Let P, Q_β be first-order stationary MC cover distribution and its corresponding stego distribution after MI embedding with parameter β. For a given $n \geq 2$, an n-element stego-system is perfectly secure iff the corresponding stego-system narrowed to 2-element cover source is perfectly secure for some $\beta_0 > 0$:*

$$\exists \beta_0 > 0,\ \forall y_1^2 \in \mathcal{X}^2 \quad P^{(2)}(X_1^2 = y_1^2) = Q_{\beta_0}^{(2)}(X_1^2 = y_1^2). \tag{2.6}$$

Moreover, the FI condition for Markov sources simplifies to

$$\forall y_1^2 \in \mathcal{X}^2 \quad \Big(P^{(2)}(X_1^2 = y_1^2) > 0\Big) \Rightarrow \Big(\frac{d}{d\beta} Q_\beta^{(2)}(y_1^2)\big|_{\beta=0} = 0\Big). \tag{2.7}$$

Proof. Because invariant distributions do not depend on β, Equation (2.6) must be valid for all $\beta > 0$ once it holds for some β_0 (see the arguments at the end of Sec. 2.1). By Corollary 6, if the stego-system is perfectly secure ($n \geq 2$), then the cover source is i.i.d. with some invariant distribution w.r.t. MI embedding and thus (2.6) and (2.7) hold. On the other hand, if (2.6) and (2.7) hold for $n = 2$ and stationary cover source, then this cover source is i.i.d. with one of k invariant distributions. This completes the proof since 2-element marginal is sufficient statistics for a first-order stationary MC. □

2.5. **Discussion.** Most practical stego-systems for digital media embed messages by making independent changes to individual cover elements. If the embedding operation is fixed, one may inquire in which cover sources the embedding is statistically undetectable in Cachin's sense. The main contribution of this part of the report is a complete geometric characterization of such sources. Using the theory of ergodic classes, it was shown that all cover sources that are perfectly secure with respect to mutually independent embedding form a vector space spanned by invariant distributions determined by the embedding operation.

Additionally, it was shown that perfect security of stegosystems with mutually independent embedding is completely captured using Fisher information formulated in Section 2.3 as the FI condition. This result not only provides a simpler and equivalent condition for perfect security, but it finds further applications in steganalysis. For example, Fisher information could be used for benchmarking such stego-systems, a direction pursued in Section 2.3. Moreover, Fisher information provides fundamental lower bounds on the variance of unbiased estimators of the change rate, which connects our results to problems in quantitative steganalysis. Finally, the FI condition plays a key role in proving the square root law of steganographic capacity of imperfect stego-systems [27, 54] (Section 1.2).

3. Experimental verification of the square root law

The PI conducted extensive experiments to validate the square root law described in the two sections above. Payloads will be embedded, using a number of different embedding methods and various payload lengths, into cover images of different sizes. Then state-of-the-art steganalysis methods will be applied to these images while looking for a square root relationship. The investigation is first directed to spatial-domain embedding and then to DCT-domain steganography for which there are some additional challenges.

3.1. **Spatial domain.** There are two difficulties to overcome in testing the theoretical result of Section 1. The first is the caveat that capacity is a square root law *all other things being equal.* Other literature on the benchmarking of steganalysis [5, 7] has shown that there are cover properties other than size – local variance, saturation, prior image processing operations – which significantly affect the detectability of payload, and it is not possible to control or even determine them all. Therefore one cannot use sets of differently-sized covers from different sources to estimate how capacity depends on size: variations in the other properties may invalidate the results. Neither can one generate small cover images by downsampling large ones, because downsampled images have a higher semantic density so, usually, higher local variance. The solution is to use a single set of large covers and repeatedly crop down to smaller images. In an attempt to preserve other image characteristics, the cropped region can be chosen so that the average local variance (here measured by average absolute difference between neighbouring pixels) is as close as possible to that of the whole image. The image libraries available to the PI are not large enough to partition them into disjoint sets for cropping to different sizes, so one may observe correlation between the content of the different-sized cropped images, but this is not expected to cause significant effects in the experiments.

The second difficulty is to define "capacity." One can set a level of detection risk which the steganographer is prepared to accept, but (even apart from the fact that the level itself will be arbitrary) how to measure detectability? As discussed in [52] and [68], there are many different detection metrics found in the literature. For these experiments the following three metrics will be considered, one standard and one very recent:

(1) The minimum sum of false positive and false negative errors for a binary classifier $P_E = \frac{1}{2}\min(P_{FA} + P_{MD})$ (for comparability with other measures, $1 - P_E$ is used);
(2) Directly from the observed cover and stego distributions of steganalysis features, a recently-developed measure called Maximum Mean Discrepancy (MMD). Its key features are described in Section 3.4.

In each case, higher values denote lower security.

Before testing this hypothesis empirically, the PI returns briefly to the definition of capacity. It is not quite correct to speak of capacity as a bound on the size of payload because it is not payload itself which is detected by steganalysis. It is the changes induced by embedding which are detected, and capacity is more properly given by a bound on permissable changes; in simple embedding schemes where the changes are of fixed magnitude, it is the number of changes one should measure. This difference is important because of the existence of adaptive source codes [30], which can exploit freedom of choice of embedding locations to reduce the number of changes required.

The first series of experiments was performed on never-compressed cover images. A set of 3000 images was downloaded from the NRCS website [63]: apparently scanned from film in full colour, these images vary slightly in size around approximately 2100×1500 pixels. The images were downsampled to a larger side of 1024 pixels, and reduced to grayscale: the same set of images has been used by a number of steganalysis researchers. Nine sets each of 3000 grayscale cover images were then created by repeated cropping, selecting the crop region best to match the local variance of the original, to sizes 100×75, 200×150, ..., 900×675.

Random payload was embedded using simple LSB replacement (for payload smaller than maximum a random selection of embedding locations was used). Three different strategies were selected for choosing the payload size according to cover size: embedding a fixed-size payload in all cover sets, embedding payload proportional to the square root of the number of cover pixels, and embedding payload proportional to the number of cover pixels. For each option, three different constants of proportionality were tested.

The method in [53] gives the currently-known best steganalysis of LSB replacement in never-compressed images, and it was applied to each set of covers and stego images. The accuracies of the resulting detector for payload, as measured by P_E and MMD, are displayed in Fig. 3.1, along with 90 % confidence intervals obtained using a simple resampling bootstrap. These experiments are in line with the theoretical predictions: whichever detectability metric is used, fixed-length payload becomes harder to detect in larger covers, fixed-proportion payload becomes easier to detect, and payload proportional to square root of cover size is (approximately) of constant detectability. At least these results suggest that square root capacity is much more plausible than proportionate capacity.

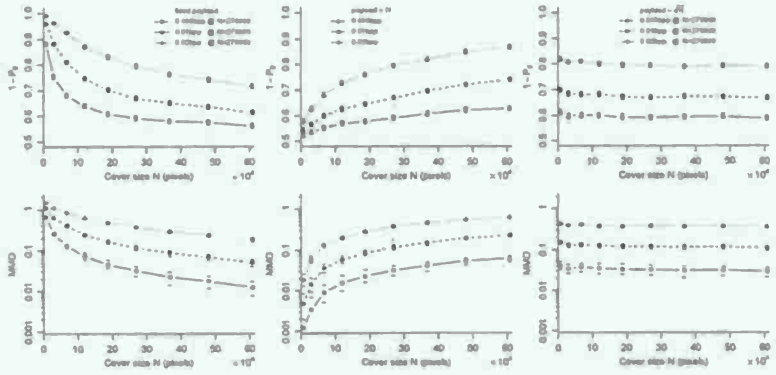

FIGURE 3.1. Detectability (y-axis: $1 - P_E$ and MMD on a log scale) as a function of cover size N (x-axis) and payload size. 90% bootstrapped confidence intervals are indicated. Left, fixed payload size. Middle, payload proportional to $\sqrt{N}$. Right, proportional to N. LSB replacement steganography in never-compressed cover images, detected by method of [47].

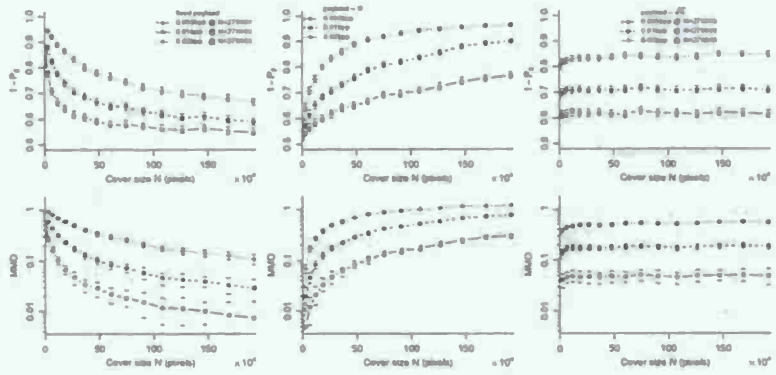

FIGURE 3.2. Detectability as a function of cover size, cf. Fig. 3.1. LSB replacement steganography in previously JPEG-compressed digital camera images, detected by method of [48].

The same experiments were repeated with a set of 1600 images taken by a Minolta DiMAGE A1 camera in raw format at a resolution of 2000×1500, subsequently converted to grayscale and subject to JPEG compression (quality factor 80). The images were cropped to 16 different sizes between 100×75 and full size, again selecting the crop region to match the average local variance of the original. When cover images have been previously compressed, different detectors for LSB replacement have better performance than that in [53], so the PI used the *Triples* detector of [48].

Charts analogous to those in Fig. 3.1, for the compressed cover images and Triples steganalysis, are displayed in Fig. 3.2 and one can draw similar conclusions: secure payload is certainly not constant, nor proportional to cover size, but appears to be approximately proportional to the square root of the cover size. More visible in this second set of experiments are artefacts in the charts for very small cover sizes, but these are to be expected if the theoretical results are only asymptotic for large covers.

Finally, the PI tested an alternative method of spatial-domain LSB embedding known as LSB matching, or ± 1 embedding. It does not have the structural flaws of LSB replacement, and seems much more difficult to detect. For the detector, the method known as the *adjacency HCF COM* found in [49] was used. This detector is still quite weak: payloads as small as those in the previous two experiments are undetectable, so the PI had to increase the payload sizes considerably. As a result, it was not possible to fit the payloads into very small covers (one cannot embed more than 1 bit per pixel using LSBs). The same 3000 never-compressed scanned images were used for the first experiment, cropped down to ten sizes between 360×270 and 900×675. The resulting charts are displayed in Fig. 3.3.

Observe that the detector's performance remains very low: P_E not much below 0.5 (which corresponds to a random detector) and MMD is near to zero (corresponding to identical distributions of cover and stego features) and, because one is digging in the detector noise, the bootstrap confidence intervals are wider. However, similar features are still apparent: falling detectability in larger covers when the payload is fixed and rising detectability when the payload is proportional to

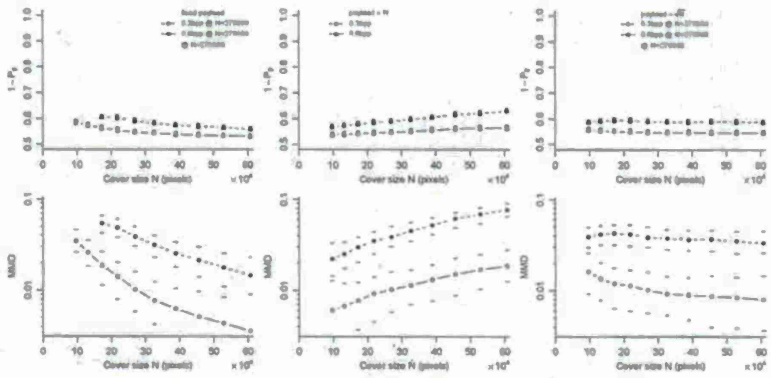

FIGURE 3.3. Detectability as a function of cover size, cf. Fig. 3.1. LSB matching steganography in never-compressed scanned images, detected by method of [49].

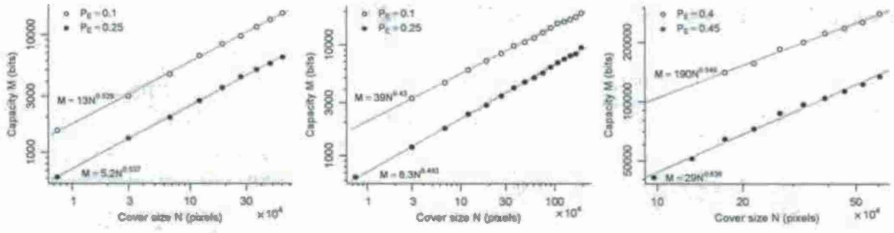

FIGURE 3.4. Capacity (y-axes, determined by limit on P_E) as a function of cover size (x-axis), log-log scale, with best-fit trend lines. Three different steganography/steganalysis methods displayed. Left, LSB replacement in never-compressed images detected by [53]; middle, LSB replacement in previously JPEG-compressed images detected by [48]; right, LSB matching detected by [49].

cover size. When the payload is proportional to the square root of the cover size, the detection metrics are *approximately* constant, although there is a suggestion that the detectability may be gradually decreasing.

To investigate more precisely how capacity depends on cover size the PI performed additional experiments: fixing on just one detection metric a bound was set on the risk to the steganographer (a minimum value of P_E) and determined the largest payload for which the detection bound can be met. This was accomplished by embedding 100 different payload sizes in each of the cover sets, measuring P_E for each combination and using linear interpolation to estimate P_E for intermediate payloads. Denoting cover size (pixels) by N and capacity (payload bits) by M, one can plot M against N on a log-log scale: if there is a relationship of the form $M \propto N^e$ then the points should fall in a straight line with slope e.

Fig. 3.4 displays the results for each of the three detectors and cover sets in the experiments, with two different thresholds for P_E (in the case of LSB matching, one must set a very high threshold for P_E because the detector is so weak). In each case, a straight line fit is determined by simple linear regression. When capacity is measured in this way, it does indeed appear to follow a relationship $M \propto N^e$, with values of e very close to 0.5. Even the line corresponding to $P_E = 0.45$ with the LSB matching detector would have slope close to 0.5 if the data points from the smallest image sets were discounted. Unfortunately one cannot use the standard least-squares tests for whether e differs *significantly* from 0.5, because the data points are not independent (they arise from images with overlapping content).

3.2. **Experimental Investigation: JPEG Steganography.** The experiments of the previous section were repeated for steganography and steganalysis in JPEG images, to see whether the square root law still holds. An improved version of F5, the so-called no-shrinkage F5 (nsF5) [37] was used as an example of a leading steganographic method in the JPEG domain. The nsF5 has the same embedding operation but uses wet paper codes [34] to remove shrinkage. The syndrome-coding mechanism in nsF5 was disabled because it introduces non-linearity between the payload and the number of embedding changes [37].

Measuring the *size* of a JPEG image is not as simple as counting pixels. After lossy compression, many of the DCT coefficients become zero and do not convey content: these coefficients cannot be used for embedding. Therefore one should define the steganographic size as the total number of nonzero DCT coefficients (abbreviated nc). This is a generally-accepted measure, although some authors also discount DC coefficients.

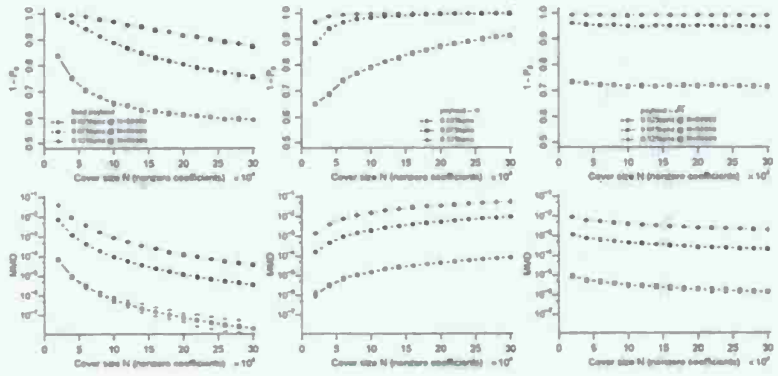

FIGURE 3.5. Detectability as a function of cover size (nonzero DCT coefficients). No-shrinkage F5 steganography with matrix embedding disabled, in JPEG covers, detected by method of [67].

The PI began with approximately 9200 never-compressed images of different sizes, and from them cropped 15 sets of cover images each with a specified number of nonzero coefficients, $2 \cdot 10^4$, $4 \cdot 10^4$, ..., $30 \cdot 10^4$, all under JPEG compression with quality factor 80. None of the images were double-compressed. As in the spatial-domain experiments, cropping was favored over scaling: the latter produces images with a higher number of nonzero DCT coefficients on higher frequencies, so statistics of DCT coefficients in scaled images vary substantially with cover size. Also paralleling the experiments in the previous section, the crop region was chosen to preserve some other characteristics of the cover. In the case of JPEG images, an attempt was made to preserve the proportion of nonzero DCT coefficients.

In each set of covers, a random message was embedded using the nsF5 algorithm. As before, the strategies for choosing the payload were to embed a fixed size payload into all cover sets, to embed payload proportional to the square root of the number of nonzero coefficients, and to embed payload proportionally to the number of nonzero coefficients.

The combination of Support Vector Machine (SVM) classifiers [16] with a Gaussian kernel and the 274-dimensional *merged feature set* [67] is the state of art general purpose steganalytic system for JPEG images. The PI measured detectability using SVMs trained specifically to each combination of cover and payload size: for each such combination, 6000 images were selected at random from the available set of 9200, split into disjoint sets of 3500 for training and 2500 for testing. In the training stage, the 3500 cover images and 3500 corresponding stego images were used; similarly in the testing stage, the 2500 cover images and 2500 corresponding stego images were all classified by the SVM. The training and testing of the SVM classifiers was repeated 100 times with different random selections of training and testing sets, and the overall $1 - P_E$ metric computed for the resulting binary classifiers.

Additionally, the MMD between the "merged feature set" vectors in cover and stego images was computed. Again, 6000 images were selected at random, this time partitioned into disjoint sets of 3000 covers and 3000 stego images (disjoint sets are necessary for good MMD estimation, see Section3.4). This was repeated 100 times with random allocations of cover and stego images: increasing the accuracy of the estimate, and also allowing the PI to estimate rough bootstrap confidence intervals. Prior to computing MMD, the vectors were normalized so that each cover feature had zero mean and unit variance: note that, although the MMD kernel γ parameter (see Section 3.4) is fixed for all cover sizes, the normalization parameters are determined separately for each set. This proved necessary because great variability was observed in the raw feature distributions, as the cover size varied.

The results of the experiment (Figure 3.5) confirm the theoretical predictions, and are similar to the results presented for the spatial domain. For fixed (respectively, linear) payload, by any metric the detectability increases (resp. decreases) with the cover size, and for payload proportional to the square root of nc the detectability is approximately constant, albeit with a barely-visible downwards trend. It is not known why the MMD measure shows this as a slightly stronger effect than P_E.

Following the previous section, the next experiment was to find payload such that the probability of error P_E matches a certain level. The search for the payload was carried under the reasonable assumption that the detectability increases with the payload size. The P_E measure at each given payload was estimated by the accuracy of the classifier (again, a SVM with a Gaussian kernel employing "merged" features) targeted to a given combination of nc and payload. The training and testing conditions were the same as in the previous experiment. Even though repeated training of the classifier is very time consuming, this approach was favoured because it provides good estimates of P_E.

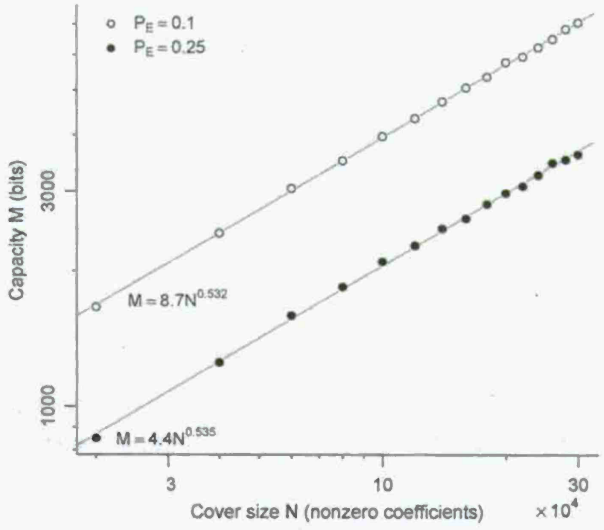

FIGURE 3.6. Capacity (y-axes, determined by limit on P_E) as a function of cover size (x-axis), log-log scale, with best-fit trend lines. No-shrinkage F5 in JPEG covers.

Figure 3.6 shows maximum payload M plotted against nc N in log-log scale for $P_E = 0.1$ and $P_E = 0.25$. Payloads were identified within 1% accuracy of the desired P_E level. In both cases, the graph shows a close accordance with a straight line and the slope of the line is close to 0.5. This shows that the capacity of JPEG images for nsF5 (without matrix embedding) grows with square root of the number of nonzero DCT coefficients.

3.3. **Discussion.** The purpose of this work was to verify the square root law of secure steganographic payload. Using carefully-designed experiments, which as far as possible isolate the effect of cover size from other cover properties, the square root law was tested for a number of steganography schemes, using contemporary steganalysis detectors. Close adherence to the law was observed.

It is not widely known that the secure capacity of a cover is proportional only to the square root of its size (where size should be measured by available embedding locations), in the absence of perfect steganography. It seems to be of fundamental importance to the practice of steganography, and could be particularly vital for the design of steganographic file systems, where the user might expect to be given an indication of secure capacity.

However, when interpreting the square root law one must take care not to ignore other important factors which contribute to capacity. In practice, properties of cover images such as saturation, local variance, and prior JPEG compression or image processing operations have been shown to have significant effects on detectability of payload [5]. One cannot simply conclude that, because one cover is twice as larger as another, it can carry $\sqrt{2}$ times the payload at an equivalent risk. The law applies *other all things being equal* and, as the difficulties constructing suitable experiments to test the law illustrate, rarely are cover images equal.

It should also be emphasised that the law truly applies not to raw payload size but to the embedding changes caused. In some embedding schemes these quantities are not proportional. For example, using syndrome coding [30] and binary embedding operations it is possible to design embedding codes for which the number of embedding changes c and payload size M approaches asymptotically the bound $c \geq Nh^{-1}(M/N)$, where h is the binary entropy function. The consequence of an asymptotic limit $c = O(\sqrt{N})$ is then $M = O(\sqrt{N}\log N)$. It would appear that, fundamentally, steganographic payload capacity is of order $\sqrt{N}\log N$.

One could argue that, because of the square root law, researchers should cease to report payloads measured in bits per pixel, bits per second, bits per nonzero coefficient, etc: the correct units should perhaps be bits per square root pixel and so on. However, such a change would still not allow comparability of different authors' benchmarks, because of the other factors affecting detectability; unless different authors use covers from the same source, their results cannot be exactly comparable in any case.

3.4. **The MMD Measure.** Maximum Mean Discrepancy (MMD) [42] is a recently-developed measure of difference between probability distributions. If X and Y are random variables with the same domain $\mathcal{X}$ then their MMD is defined

as

$$\max \left| \mathbf{E}[f(X)] - \mathbf{E}[f(Y)] \right|, \tag{3.1}$$

where the maximum is taken over all mappings $f : \mathcal{X} \mapsto \mathbb{R}$ from a unit ball $\mathcal{F}$ in a Reproducing Kernel Hilbert Space (RKHS). Although not a true metric, the MMD is symmetric, nonnegative, and zero only when X and Y have the same distribution.

For technical reasons it is simpler to use the square of the MMD measure in (3.1) and in this report the PI always reports squared MMD values. Given n independent observations $\mathbf{x} = (x_1, \ldots, x_n)$ of X and a further n independent observations $\mathbf{y} = (y_1, \ldots, y_n)$ of Y, the (squared) MMD may be estimated by

$$\frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + k(y_i, y_j) - 2k(x_i, y_j)$$

where k is a bounded universal kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ that defines the dot product in the RKHS [82]. The variance of the estimator decreases as $1/\sqrt{n}$, almost independently of the dimension of the random variables [43], and can also be improved by bootstrapping. MMD has been used for comparing security of stego-schemes in [68].

In this report, MMD is measured w.r.t. a Gaussian kernel

$$k(x, y) = \exp(-\gamma \|x - y\|^2),$$

with the width parameter γ set to η^{-2}, where η is the median of the L_2-distances between (normalized) features in a pooled set of all cover images. This choice is justified in [68]. Note that, for direct comparison of MMD values obtained from experiments on different cover sets, the γ parameter should remain fixed.

4. Fisher Information Determines Capacity of ε-secure Steganography

As explained in Section 1, most practical stegosystems for digital media work by applying a mutually independent embedding operation to each element of the cover. For such stegosystems, the PI has shown in Section 2 that the Fisher information w.r.t. the change rate is a perfect security descriptor equivalent to KL divergence between cover and stego images. Under the assumption that cover elements form a Markov chain, in this section a closed-form expression for the Fisher information is derived and it is also shown how it can be used for comparing stegosystems and optimizing their performance. In particular, using an analytic cover model fit to experimental data obtained from a large number of natural images, the PI proves that the ± 1 embedding operation is asymptotically optimal among all mutually independent embedding operations that modify cover elements by at most 1.

The key concept in essentially all communication schemes is the channel capacity defined as the amount of information, or largest payload, that can be safely transmitted over the channel. As shown in Section 1, for imperfect stegosystems, the communication *rate* is not a good descriptor of the channel because it approaches zero with increasing n. The sender, however, still needs to know what level of risk she is exposing herself to when sending a message to the recipient. It is critical for the sender to know how much information she can send using her stegosystem in an n-element cover while keeping the KL divergence between cover and stego objects below some chosen ε. The SRL informs the sender that the amount of information that she can hide scales as $r\sqrt{n}$ [27], with r constant.

In this section, the proportionality constant r from the SRL is proposed as a more refined measure of steganographic capacity of imperfect stegosystems. By the form of the law, r, for which the PI coins the term *the root rate*, essentially expresses the capacity per square root of cover size. Under the assumption that covers form a Markov chain and embedding is realized by applying a sequence of independent embedding operations to individual cover elements, a closed form expression is derived for the root rate. The root rate depends on the Fisher information rate w.r.t. the the change rate, which is a perfect security descriptor equivalent to the KL divergence between distributions of cover and stego objects (Section 2). Expressing the Fisher information rate analytically as a quadratic form allows one to evaluate, compare, and optimize security of stegosystems. To this end, the PI derives an analytic cover model from a large database of natural images represented in the spatial domain and shows that the ± 1 embedding operation is asymptotically optimal among all mutually independent embedding operations that modify cover elements by at most 1. Finally, using the Fisher information rate, the PI compares security of several practical stegosystems, including LSB embedding and ± 1 embedding. The findings appear to be consistent with results previously obtained experimentally using steganalyzers and are in good agreement with the experimental study of Section 3.

The notation and definitions carry over from Sections 1 and 2.

4.1. Fisher information in steganography.

The PI now introduces the concept of root rate as a measure of capacity of imperfect stegosystems. First, the relationship between steganographic capacity of stegosystems satisfying Assumptions 1–3 and the Fisher information w.r.t. the parameter β

$$I_n(0) = E_P\left[\left(\frac{d}{d\beta}\ln Q_\beta^{(n)}(y_1^n)\Big|_{\beta=0}\right)^2\right] \tag{4.1}$$

is explained. Then in Section 4.2, a closed form expression for the Fisher information is derived and written in terms of the expected relative payload α instead of parameter β as this form is more informative for the steganographer.

Fisher information is a fundamental quantity that frequently appears in theoretical steganography and in general in signal detection and estimation. For example, the Cramer-Rao lower bound states that the reciprocal of Fisher information, $1/I_n(\beta)$, is the lower bound on the variance of unbiased estimators of β (quantitative steganalyzers). Fisher information also appears in the leading term of Taylor expansion of the KL divergence $d_n(\beta) \triangleq D_{KL}(P^{(n)}||Q_\beta^{(n)}) = \beta^2 I_n(0)/(2\ln 2)+O(\beta^3)$, where

$$D_{KL}\left(P^{(n)}||Q_\beta^{(n)}\right) \triangleq \sum_{x_1^n\in\mathcal{X}^n} P^{(n)}(x_1^n)\log_2\frac{P^{(n)}(x_1^n)}{Q_\beta^{(n)}(x_1^n)}.$$

Zero KL divergence implies zero Fisher information. Although the opposite is not true in general, it holds for all stegosystems with MI embedding and arbitrary cover model. For such stegosystems, Fisher information $I_n(0)$ represents a perfect security descriptor equivalent to the KL divergence. Fisher information was also proposed for benchmarking steganalyzers [52].

The relationship between the Fisher information rate and steganographic capacity of stegosystems satisfying Assumptions 1–3 was established in Section 2. It was essentially shown that such stegosystems are subject to the Square Root Law, which means that payloads that grow faster than $\sqrt{n}$, i.e., $\lim_{n\to\infty}\beta(n)n/\sqrt{n}=\infty$, can be detected arbitrarily accurately,

whereas payloads that grow slower than $\sqrt{n}$, i.e., $\beta(n)n/\sqrt{n} \leq K < \infty$, lead to ε-secure stegosystems, $d_n(\beta) < \varepsilon$.[3] This result means that the payload that can be securely transmitted over the steganographic channel scales as $r\sqrt{n}$. Consequently, the sequence of embedding parameters $\beta(n)$ must approach zero for ε-secure systems and thus the communication rate tends to zero. Due to this fact, it makes sense to evaluate steganographic capacity in the limit of $\beta(n) \to 0$.

4.2. **Root rate.** The problem of steganalysis can be formulated as the following hypothesis testing problem

$$\begin{aligned} \mathrm{H}_0 &: \beta = 0 \\ \mathrm{H}_1 &: \beta > 0. \end{aligned} \tag{4.2}$$

It is shown show that for small (and known) β and large n, the likelihood ratio test with test statistic

$$\frac{1}{\sqrt{n}} T^{(n)}_{\beta_0}(X) = \frac{1}{\sqrt{n}} \ln \left(Q^{(n)}_{\beta_0}(X) / P^{(n)}(X) \right), \tag{4.3}$$

is a mean-shifted Gauss-Gauss problem.[4] This property, usually called the Local Asymptotic Normality (LAN) of the detector, allows us to quantify and correctly compare security of embedding algorithms operating on the same MC cover model for small values of β.

In this case, the detector performance can be completely described by the deflection coefficient d^2, which parametrizes the ROC curve as it binds the probability of detection, P_D, as a function of the false alarm probability, P_{FA},

$$P_D = Q\big(Q^{-1}(P_{FA}) - \sqrt{d^2}\big).$$

Here, $Q(x) = 1 - \Phi(x)$ and $\Phi(x)$ is the cdf of a standard normal variable $N(0,1)$. Large value of the deflection coefficient implies better detection or weaker steganography.

First, the PI states the LAN property for the HMC model w.r.t. the embedding parameter β and then extends this result with respect to the relative payload α.

Theorem 8. [LAN of the LLRT] *Under Assumptions 1–3, the likelihood ratio* (4.3) *satisfies the local asymptotic normality (LAN), i.e., under both hypotheses and for values of β up to order β^2*

$$\sqrt{n}\big(T^{(n)}_\beta/n + \beta^2 I/2\big) \xrightarrow{d} N(0, \beta^2 I) \text{ under } \mathrm{H}_0 \tag{4.4}$$

$$\sqrt{n}\big(T^{(n)}_\beta/n - \beta^2 I/2\big) \xrightarrow{d} N(0, \beta^2 I) \text{ under } \mathrm{H}_1, \tag{4.5}$$

where I is the Fisher information rate, $I = \lim_{n\to\infty} \frac{1}{n} I_n(0)$, and $\xrightarrow{d}$ is the convergence in distribution. The detection performance is thus completely described by the deflection coefficient

$$d^2 = \frac{(\sqrt{n}\beta^2 I/2 + \sqrt{n}\beta^2 I/2)^2}{\beta^2 I} = n\beta^2 I.$$

Proof. Due to limited space, only a brief outline of the proof is provided. The Gaussianity of the test statistic follows from the Central Limit Theorem (CLT) due to the fact that the test statistic is close to being i.i.d. Formal proof of this uses exponential forgetting of the prediction filter [18, Lemma 9] and follows similar steps as the proof of the CLT for Markov chains [17]. The mean and variance of the likelihood ratio (4.3) is obtained by expanding (4.3) in Taylor series w.r.t. β and realizing that the leading term is the quadratic term containing the Fisher information rate. □

The conclusion of the theorem is now rephrased in terms of the payload rather than the parameter β. Matrix embedding (syndrome coding) employed by the stegosystem may introduce a non-linear relationship $\beta = f(\alpha)$ between both quantities. In general, the payload embedded at each cover element may depend on its state $i \in \mathcal{X}$. Thus, the expected value of the relative payload that can be embedded in each cover is $\alpha(\beta) = \sum_{i\in\mathcal{X}} \pi_i \alpha_i(\beta)$, where $\alpha_i(\beta)$ stands for the number of bits that can be embedded into state $i \in \mathcal{X}$ and π_i is the stationary distribution of the MC. The value of β for which α is maximal will be denoted as β_{MAX}

$$\beta_{MAX} = \arg\max_\beta \alpha(\beta).$$

For example, for ternary ± 1 embedding $\beta_{MAX} = 2/3$ and $\alpha_i(\beta_{MAX}) = \log_2 3$, while for binary ± 1 embedding $\beta_{MAX} = 1/2$ and $\alpha_i(\beta_{MAX}) = 1$. The matrix $\mathbb{C}$ is the same for both embedding methods. The only formal difference is the range of the

[3] Here, it is assumed that there exists a linear relationship between $\beta(n)$ and the relative payload $\alpha(n)$ (e.g., the stegosystem does not employ matrix embedding). Indeed, application of matrix embedding does not invalidate our arguments as $\alpha(n)$ differs from $\beta(n)$ only by a multiplicative factor bounded by $\log n$.

[4] In hypothesis testing, the problem of testing $N(\mu_0, \sigma^2)$ vs. $N(\mu_1, \sigma^2)$ is called the mean-shifted Gauss-Gauss problem and its detection performance is completely described by the deflection coefficient $d^2 = (\mu_0 - \mu_1)^2/\sigma^2$ [46, Chapter 3].

parameter β. It is also worth noting that unless all α_i are the same, the maximal payload will depend on the distribution of individual states π_i.

To simplify our arguments, it will be assumed that a linear relationship between β and α exists (e.g., matrix embedding is not considered). Therefore, one can write

$$\beta = f(\alpha) = \frac{\beta_{MAX}}{\alpha_{MAX}}\alpha, \tag{4.6}$$

where $\alpha \in [0, \alpha_{MAX}]$ and $\alpha_{MAX} = \alpha(\beta_{MAX})$ denotes the average number of bits that can be embedded into cover element while embedding with $\beta = \beta_{MAX}$ (maximum change rate).

From (4.6), the deflection coefficient can be expressed in terms of the relative payload α by substituting $\beta = f(\alpha)$ from (4.6) into Q_β

$$d^2 = n\alpha^2\left(\frac{\beta_{MAX}}{\alpha_{MAX}}\right)^2 I. \tag{4.7}$$

In practice, the sender can control statistical detectability by bounding $d^2 < \varepsilon$ for some fixed ε, obtaining thus an upper bound on the total number of bits (payload) αn that can be safely embedded (this requires rearranging the terms in (4.7))

$$\alpha n \le \frac{\alpha_{MAX}}{\beta_{MAX}}\sqrt{\frac{\varepsilon}{I}}n. \tag{4.8}$$

In analogy to the communication rate, it is natural to define *the root rate*

$$r \triangleq \frac{\alpha_{MAX}}{\sqrt{I}\beta_{MAX}} \tag{4.9}$$

as the quantity that measures steganographic security of imperfect stegosystems in bits per square root of cover size per square root of KL divergence. The root rate can be used for comparing stegosystems with a MC cover model.

In the next theorem, proved in the appendix, the PI establishes the existence of the main component of the root rate, the Fisher information rate I, and expresses it in a closed form.

Theorem 9. [Fisher information rate] *Let $\mathbb{A} = (a_{ij})$ define the MC cover model and $\mathbb{B}$, defined by matrix $\mathbb{C} = (c_{ij})$, capture the embedding algorithm. Then, the normalized Fisher information $I_n(0)/n$ approaches a finite limit I as $n \to \infty$. This limit can be written as $I = \mathbf{c}^T\mathbb{F}\mathbf{c}$, where $\mathbf{c}$ is obtained by arranging $\mathbb{C}$ into a column vector of size N^2 with elements c_{ij}.[5] The matrix $\mathbb{F}$ of size $N^2 \times N^2$ is defined only in terms of matrix $\mathbb{A}$ and does not depend on the embedding algorithm. The elements of matrix $\mathbb{F}$ are*

$$f_{(i,j),(k,l)} = [j = l]V(i,j,k) - U(i,j,k,l), \tag{4.10}$$

where by the Iverson notation $[j = l]$ is one if $j = l$ and zero otherwise and

$$V(i,j,k) = \left(\sum_{z\in\mathcal{X}} \pi_z a_{zi}\frac{a_{zk}}{a_{zj}}\right)\left(\sum_{z\in\mathcal{X}} a_{iz}\frac{a_{kz}}{a_{jz}}\right)$$

$$U(i,j,k,l) = \pi_i\left(a_{ik} - a_{il}\frac{a_{jk}}{a_{jl}}\right) + \pi_k\left(a_{ki} - a_{kj}\frac{a_{li}}{a_{lj}}\right).$$

Moreover, $|I_n(0)/n - I| \le C/n$ for some constant C. This constant depends only on the elements of matrix $\mathbb{A}$ and not on the embedding algorithm. The quadratic form $I(\mathbf{c}) = \mathbf{c}^T\mathbb{F}\mathbf{c}$ is semidefinite, in general.

By inspecting the proof of the theorem, the matrix $\mathbb{F}$ can be seen as the Fisher information rate matrix w.r.t. the parameters $\{b_{ij}|1 \le i,j \le N\}$. It describes the natural sensitivity of the cover source to MI embedding. The quadratic form then combines these sensitivities with coefficients given by the specific embedding method and allows us to decompose the intrinsic detectability caused by the cover source from the detectability caused by the embedding algorithm.

Corollary 10. *For the special case when the MC degenerates to an i.i.d. cover source with distribution $P = \pi$, the Fisher information rate simplifies to*

$$I = \sum_{i,j,k\in\mathcal{X}} c_{ij}\frac{\pi_i\pi_k}{\pi_j}c_{kj}.$$

[5] The order of elements in $\mathbf{c}$ is immaterial as far as the same ordering is used for pairs (i,j) and (k,l) in matrix $\mathbb{F}$.

4.3. **Maximizing the root rate.** In the previous section, it was established that the steganographic capacity of imperfect stegosystems should be measured as the root rate (4.9) defined as the payload per square root of the cover size and per square root of KL divergence. The most important component of the root rate is the stegosystem's Fisher information rate, for which an analytic form was derived in Theorem 9. The steganographer is interested in designing stegosystems (finding $\mathbb{C}$) with the highest possible root rate. This can be achieved by minimizing the Fisher information rate or by embedding symbols from a larger alphabet, i.e., increasing the ratio α_{MAX}/β_{MAX}. In this section, two general strategies are described for maximizing the root rate that are applicable to practical stegosystems. In Section 4.6, conclusions are drawn from experiments when these strategies are applied to real cover sources formed by digital images.

Before proceedings with further arguments, the PI points out that the highest root rate is obviously obtained when the Fisher information rate is zero, $I = 0$. This can happen for non-trivial embedding ($\mathbb{C} \neq 0$) in certain sources because the Fisher information rate is a semidefinite quadratic form. Such stegosystems, however, would be perfectly secure and thus by Assumption 3 are excluded from our consideration.[6]

The number of bits, α_i, that can be embedded at each state $i \in \mathcal{X}$ is bounded by the entropy of the ith row of $\mathbb{B} = \mathbb{I} + \beta\mathbb{C}$, $H(\mathbb{B}_{i\bullet})$. Thus, in the most general setting, one wishes to maximize the root rate

$$\frac{\sum \pi_i H\left(\mathbb{B}_{i\bullet}(\beta_{MAX})\right)}{\beta_{MAX}} \frac{1}{\sqrt{I}}$$

w.r.t. matrix $\mathbb{C}$. The nonlinear objective function makes the analysis rather complicated and the result may depend on the distribution of individual states π. Moreover, even if one knew the optimal solution, care needs to be taken in interpreting such results, because a practical algorithm allowing us to communicate the entropy of the additive noise may not be available. The PI is aware of only a few practical embedding algorithms that communicate the maximal amount of information (LSB embedding with binary symbols and ± 1 embedding with ternary symbols). In practice, stochastic modulation [32] can be used in some cases to embed information by adding noise with a specific pmf (matrix $\mathbb{C}$), but the specific algorithms described in [32] are suboptimal.

In the rest of this section, two different approaches are presented how to optimize the embedding algorithm under different settings that are practically realizable.

4.4. **Optimization by convex combination of known methods.** One simple and practical approach to optimize the embedding method is obtained by combining existing stegosystems $S^{(1)}$ and $S^{(2)}$. Suppose the sender embeds a portion of the message into λn elements, $0 < \lambda < 1$, using $S^{(1)}$ and use the remaining $(1-\lambda)n$ elements to embed the rest of the message using $S^{(2)}$. If both sender and the receiver select the elements pseudo-randomly based on a stego key, the impact on a single cover element follows a distribution obtained as a convex combination of the noise pmfs of both methods. Note that the methods are allowed to embed a different number of bits per cover element since the receiver knows which symbol to extract from each part of the stego object. Let $S^{(i)}$ represent the ith embedding method with matrix $\mathbb{C}^{(i)}$, or its vector representation $\mathbf{c}^{(i)}$, with ratio $\rho^{(i)} = \alpha^{(i)}_{MAX}/\beta^{(i)}_{MAX}$ for $i \in \{1,2\}$. The root rate $r(\lambda)$ of the method obtained by the above approach (convex embedding) with parameter λ can be written as

$$\begin{aligned} r(\lambda) &= \frac{\lambda\rho^{(1)} + (1-\lambda)\rho^{(2)}}{\sqrt{(\lambda\mathbf{c}^{(1)} + (1-\lambda)\mathbf{c}^{(2)})^T \mathbb{F}(\lambda\mathbf{c}^{(1)} + (1-\lambda)\mathbf{c}^{(2)})}} \\ &= \frac{\lambda\rho^{(1)} + (1-\lambda)\rho^{(2)}}{\sqrt{\lambda^2 I^{(1)} + (1-\lambda)^2 I^{(2)} + 2\lambda(1-\lambda)I^{(1,2)}}}, \end{aligned} \tag{4.11}$$

where $I^{(i)}$ is the Fisher information rate of $S^{(i)}$ and $I^{(1,2)} = \left(\mathbf{c}^{(1)}\right)^T \mathbb{F}\mathbf{c}^{(2)}$. Here, the symmetry of $\mathbb{F}$ was used to write $I^{(1,2)} = I^{(2,1)}$.

4.5. **Minimizing the Fisher information rate.** In an alternative setup, one may deal with the problem of optimizing the shape of the additive noise pmf under the assumption that the number of bits, α_i, embedded at each state $i \in \mathcal{X}$ is constant. For example, one may wish to determine the optimal pmf that would allow communicating 1 bit per element ($\alpha_i = 1, \forall i \in \mathcal{X}$) by changing each cover element by at most 1. In this problem, the ratio α_{MAX}/β_{MAX}, as well as the cover model (matrix $\mathbb{A}$), are fixed and known. The task is to minimize the Fisher information rate I.

The PI now formulates the optimization problem by restricting the form of the matrix $\mathbb{C} = (c_{ij})$, or its vector representation $\mathbf{c} = (c_{ij}) \in R^{N^2 \times 1}$, to the following linear parametric form

$$\mathbf{c} = \mathbb{D}v + e, \tag{4.12}$$

[6] An example of such a stegosystem is LSB embedding in i.i.d. covers with $\pi_{2i} = \pi_{2i+1}$ for all i.

where $\mathbb{D} = (d_{ij})$ is a full-rank real matrix of size $N^2 \times k$, e is a real column vector of size N^2, and $v = (v_1, \dots, v_k)^T$ is a k-dimensional column vector. It will be assumed that $v \in \mathcal{V}$, where $\mathcal{V}$ is bounded by a set of linear inequalities[7] and the constraint $\sum_j c_{ij} = 0$ for all $i \in \{1, \dots, N\}$. In other words, the matrix $\mathbb{C}$ is decomposed into k real parameters v_i, $i \in \{1, \dots, k\}$. The following example shows one such representation for a stegosystem whose embedding changes are at most 1.

Example 11. [Tridiagonal embedding] Set $c_{ii} = -1$, $c_{i,i-1} = v_{i-1}$, and $c_{i,i+1} = 1 - v_{i-1}$ for $i \in \{2, \dots, N-1\}$ (and suitably defined at the boundaries). This allows modeling ± 1 embedding, LSB embedding, and all possible MI embedding methods that modify every element by at most 1. By setting $c_{ii} = -1$ for all i, the sender constraints her choices to stegosystems that embed the same payload into every state $i \in \mathcal{X}$ for all $\beta \geq 0$. This model has $k = N - 2$ parameters and the set $\mathcal{V}$ is formed by $v_j \in [0, 1]$, $j \in \{1, \dots, k\}$.

The sender's task is to minimize the Fisher information rate for embedding methods given by (4.12). The function $I(v) = (\mathbb{D}v + e)^T \mathbb{F} (\mathbb{D}v + e)$ can attain its minimum either at a point with a zero gradient[8] (a critical point) or on the boundary of $\mathcal{V}$. The PI now derives a set of linear equations for the set of all possible critical points. This approach will be used in Section 4.6 to prove that ternary ± 1 embedding is asymptotically optimal within the class of tridiagonal embedding in spatial domain.

For the chosen parametrization, the gradient w.r.t. every parameter v_j can be expressed as

$$\frac{\partial}{\partial v_j} I(v) = \frac{\partial}{\partial v_j} (\mathbb{D}v + e)^T \mathbb{F} (\mathbb{D}v + e) = 2(\mathbb{D}_{\bullet j})^T \mathbb{F} (\mathbb{D}v + e),$$

where $\mathbb{D}_{\bullet j}$ is the jth column of matrix $\mathbb{D}$. Because every possible candidate v_0 for the optimal parameters must satisfy $(\partial / \partial v_j) I(v)|_{v=v_0} = 0$ for every $j \in \{1, \dots, k\}$, all critical points are solutions of the following linear system

$$\mathbb{D}^T \mathbb{F} \mathbb{D} v = -\mathbb{D}^T \mathbb{F} e. \tag{4.13}$$

If this system has a unique solution $v_0 \in \mathcal{V}$, then v_0 corresponds to matrix $\mathbb{C}$ achieving the global minimum of the Fisher information rate, which corresponds to the best MI embedding method w.r.t. $\mathcal{V}$ and a given MC cover source.

4.6. **Experiments.** In the previous section, two strategies for maximizing the root rate for practical stegosystems were outlined. This section presents specific results when these strategies are applied to stegosystems operating on 8-bit gray-scale images represented in the spatial domain. Although images are two dimensional objects with spatial dependencies in both directions, they are represented in a row-wise fashion as a first-order Markov Chain over $\mathcal{X} = \{0, \dots, 255\}$. The MC model represents the first and simplest step of capturing pixel dependencies while still retaining the important advantage of being analytically tractable. Then, a parametric model is adopted for the transition probability matrix of this Markov cover source and it is shownto be a good fit for the empirical transition probability matrix $\mathbf{A}$ estimated from a large number of natural images. This analytic model is used to evaluate the root rate (4.9) of several stegosystems obtained by a convex combinations of known methods. Finally, it is shown that the optimal embedding algorithm that modifies cover elements by at most 1 is very close to ± 1 embedding.

In principle, in practice one could calculate the Fisher information rate using equation (4.10) with an empirical matrix $\mathbf{A}$ estimated from a large number of images. However, this approach may give misleading results because (4.10) is quite sensitive to small perturbations of a_{ij} with a small value (observe that $I = +\infty$ if $a_{ij} = 0$). This is not going to be an issue in practice since rare transitions between distant states are probable but content dependent, which makes them difficult to be utilized for steganalysis. Because small values of a_{ij} can not be accurately estimated in practice, the matrix $\mathbf{A}$ is represented with the following parametric model

$$a_{ij} = \frac{1}{Z_i} e^{-(|i-j|/\tau)^\gamma}, \tag{4.14}$$

where $Z_i = \sum_{j=1}^{256} e^{-(|i-j|/\tau)^\gamma}$ is the normalization constant. The parameter γ controls the shape of the distribution, whereas τ controls its "width." The model parameters were found in the logarithmic domain using the least square fit between (4.14) and its empirical estimate. To validate this model, the least square was carried out fit separately for three image databases: never compressed images taken by several digital cameras[9] (CAMRAW), digital scans[10] (NRCS),

[7] E.g., one must have $\mathbb{B} \geq 0$.

[8] Note that the semidefiniteness of $\mathbb{F}$ guarantees that the extremum must be a minimum.

[9] Expanded version of CAMERA_RAW database from [41] with 4547 8-bit images.

[10] Contains 2375 raw scans of negatives coming from the USDA Natural Resources Conservation Service (`http://photogallery.nrcs.usda.gov`).

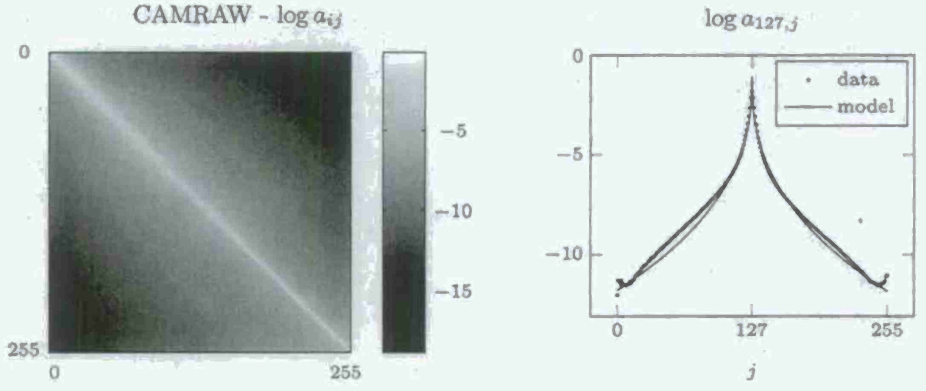

FIGURE 4.1. Left: plot of the empirical matrix A estimated from CAMRAW database in log domain. Right: comparison of the 128th row of matrix A estimated from the same database with the analytic model (4.14).

and decompressed JPEG images[11] (NRCS-JPEG). Figure 4.1 shows the comparison between the empirical matrix A estimated from the CAMRAW database and the corresponding fit. Although this model cannot capture some important macroscopic properties of natural images, such as pixel saturations, it remains analytically tractable and is valid for many natural images.

The left part of Figure 4.2 shows the root rate (4.11), $r(\lambda)$, for a convex combination of LSB and ± 1 embedding, $\lambda \in [0, 1]$, for different image sources. The higher the root rate $r(\lambda)$, the better the stegosystem. The results are consistent with the thesis that ± 1 embedding is less detectable than LSB embedding. Similarly, the capacity of stegosystems with covers from NRCS (scans) is believed to be higher than the capacity of stegosystem with decompressed JPEGs or images from digital cameras. This fact is in agreement with the obtained result for all values of the convex combination of LSB and ± 1 embedding and it can be attributed it to the fact that scans contain a higher level of noise that masks embedding changes. In contradiction with expectations, decompressed JPEGs from NRCS-JPEG have a higher root rate than raw images from digital cameras (CAMRAW). This phenomenon is probably caused by the simplicity of the MC model, which fails to capture JPEG artifacts because they span across larger distances than neighboring pixels.

The methodology described in Section 4.5 is now used to maximize the root rate with respect to stegosystems that modify each cover element by at most 1. It is done for the cover model fit obtained from the NRCS database. Assuming the embedding operation is binary, it can embed one bit per cover element. Thus, it is sufficient to find the MI embedding that attains the minimum Fisher information rate. The PI used the parametrization from Example 11 and solved the system of equations (4.13). This system has only one solution $v = (v_1, \ldots, v_{254}) \in \mathcal{V} = [0, 1]^{254}$ and thus it represents MI embedding with the minimum Fisher information rate. This solution is shown in the right part of Figure 4.2 along with the representation of the ± 1 embedding operation. The optimal MI embedding differs from ± 1 embedding only at the boundary of the dynamic range. This is due to the finite number of states in the MC model. It was experimentally verified that the relative number of states with $|v_i - 0.5| \geq \delta$ tends to zero for a range of $\delta > 0$ as $N \to \infty$ for fixed parameters of the analytic model.[12] Thus, the boundary effect is negligible for large N. This suggests that the loss in capacity when using ± 1 embedding algorithm is negligible for large N or, in other words, ± 1 embedding is asymptotically optimal.

4.7. **Discussion.** In sharp contrast with the well-established fact that the steganographic capacity of perfectly secure stegosystems increases linearly with the number of cover elements, n, the square root law states that steganographic capacity of a quite wide class of imperfect stegosystems is only proportional to $\sqrt{n}$. The communication rate of imperfect stegosystems is thus non-informative because it tends to zero with n. Instead, an appropriate measure of capacity is the constant of proportionality in front of $\sqrt{n}$, for which the term the *root rate* is coined whose unit is bit per square root of cover size per square root of KL divergence. The root rate is shown to be inversely proportional to the square root of the Fisher information rate of the stegosystem. Adopting a Markov model for the cover source, an analytic formula for the root rate can be derived with Fisher information rate expressible as a quadratic form defined by the cover transition probability matrix evaluated at a vector fully determined by the embedding operation. This analytic form is important as it enables comparing capacities of imperfect stegosystems as well as optimizing their embedding operation (maximize

[11]Images from NRCS database compressed with JPEG quality factor 70.

[12]The same is likely to be true for all $\delta > 0$.

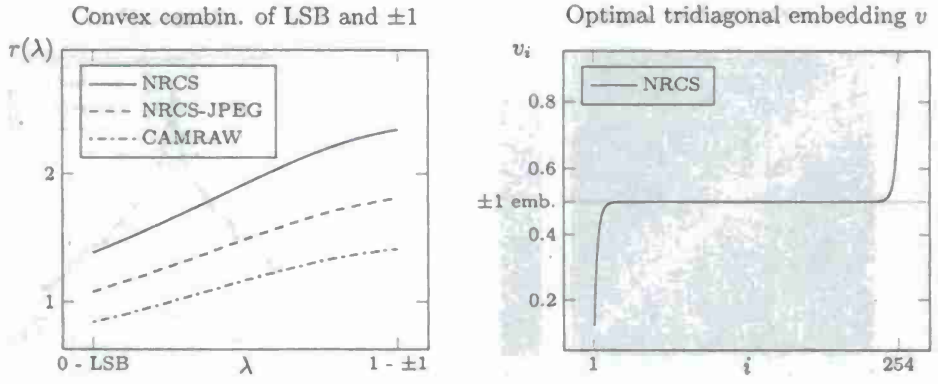

FIGURE 4.2. Left: the root rate $r(\lambda) = \alpha_{MAX}/(\beta_{MAX}\sqrt{I})$ of a convex combination of LSB and ±1 embedding for different image sources. Right: optimal parameters $v = (v_1, \ldots, v_{254})$ of MI embedding (4.12) minimizing the Fisher information rate while modifying cover elements by at most 1. The difference between ±1 embedding and optimal MI embedding is due to boundary effects that vanish as $N \to \infty$.

the root rate). By fitting a parametric model through the empirical transition probability matrix for neighboring pixels of real images, the model is used to compute and compare the root rate of known steganographic schemes and their convex combinations. In agreement with results previously established experimentally using blind steganalyzers, the analysis indicates that ternary ±1 embedding is more secure than LSB embedding and it is also optimal among all embedding methods that modify pixels by at most 1. Furthermore, by analyzing image databases of raw images from different sources, it has been established that the root rate is larger for images with higher noise level as is to be expected. Among the surprising results of our effort, the PI points out the fact that the root rate for ±1 embedding is only about twice larger than for LSB embedding, which contrasts with the fact that current best steganalyzers for LSB embedding are markedly more accurate than the best steganalyzers of ±1 embedding. This hints at the existence of significantly more accurate detectors of ±1 embedding that are yet to be found.

4.8. **Proofs. Proof of Theorem 9:** Only the main idea of the proof is presented in this report, leaving the main bulk of technical details to the report [19]. The decomposition of the sequence $I_n(0)/n$ to a quadratic form and its properties can be obtained directly from the definition of Fisher information

$$\frac{1}{n}I_n(0) = \frac{\ln 2}{n}\frac{\partial^2}{\partial\beta^2}d_n(\beta)\Big|_{\beta=0} =$$
$$= -\sum_{(i,j)}\sum_{(k,l)}\frac{\ln 2}{n\ln 2}E_P\Big[\underbrace{\Big(\frac{\partial^2}{\partial b_{ij}b_{kl}}\ln Q_\beta(Y_1^n)\Big|_{\mathbb{B}=\mathbb{I}}\Big)}_{\triangleq g(Y_1^n,i,j,k,l)}\Big]\underbrace{\Big(\frac{\partial b_{ij}}{\partial\beta}\Big|_{\beta=0}\Big)}_{=c_{ij}}\underbrace{\Big(\frac{\partial b_{kl}}{\partial\beta}\Big|_{\beta=0}\Big)}_{=c_{kl}}.$$

The derivatives of the log-likelihood are evaluated at $\mathbb{B} = \mathbb{I}$ because $\mathbb{B}(\beta) = \mathbb{I} + \beta\mathbb{C}$ and $\beta = 0$. By using $Q_\beta(y_1^n) = \sum_{x_1^n\in\mathcal{X}^n}P(x_1^n)Q_\beta(y_1^n|x_1^n)$, the random variable $g(Y_1^n, i, j, k, l)$ does not depend on the embedding method. This is because the derivatives are evaluated at $\mathbb{B} = \mathbb{I}$ and thus only contain the elements of the cover source transition matrix $\mathbf{A}$. The proof of the convergence of $-\frac{1}{n}E_P[g(Y_1^n, i, j, k, l)]$ to $f_{(i,j),(k,l)}$ and its closed form is more involved and is presented in the report [19]. The semidefinitness of the quadratic form follows from semidefiniteness of the Fisher information matrix $\mathbb{F}$. It is not positively definite because for an i.i.d. cover source all rows of matrix $\mathbb{F}$ coincide and are thus linearly dependent.

5. Constructing Practical Stegosystems using Syndrome-Trellis Codes

The previous sections describe the achievements in fundamental understanding of steganographic capacity in imperfect stegosystems. This section describes a complete practical methodology for constructing minimal-distortion steganographic systems in practice. By this, the PI means how to embed a given payload while introducing the smallest possible distortion. As long as one can tie the distortion to statsitical detectability, this approach provides (near) optimal constructions.

The idea is to proceed by steps from an easier problem to more complex ones. The simplest problem is to optimize the embedding for an additive binary embedding operation defined as follows: First, assign to each over element a scalar expressing the distortion of an embedding change done by replacing the cover element by its "other" value (here, the assumption is that the embedding operation is binary). The total distortion is assumed to be a sum of per-element distortions. Both the payload-limited sender (minimizing the total distortion while embedding a fixed payload) and the distortion-limited sender (maximizing the payload while introducing a fixed total distortion) are considered in this section. The non-binary case can be decomposed into several binary cases by replacing the individual bits in cover elements as described in one of the papers by the PI [25, 26]. The binary case can be resolved using a novel syndrome-coding scheme based on dual convolutional codes equipped with the Viterbi algorithm. This fast and very versatile solution achieves state-of-the-art results in steganographic applications while having linear time and space complexity w.r.t. the number of cover elements. The PI illustrates the power of the constructions for various relative payloads and different distortion profiles, including the wet paper channel. This framework substantially improves upon current coding schemes used in steganography, such as matrix embedding and wet paper codes.

There exist two mainstream approaches to steganography in empirical covers, such as digital media objects: steganography designed to preserve a chosen cover model and steganography minimizing a heuristically-defined embedding distortion. The strong argument for the former strategy is that provable undetectability can be achieved w.r.t. a specific model. The disadvantage is that an adversary can usually rather easily identify statistical quantities that go beyond the chosen model that allow reliable detection of embedding changes. The latter strategy is more pragmatic – it abandons modeling the cover source and instead tells the steganographer to embed payload while minimizing a distortion function. In doing so, it gives up any ambitions for perfect security. Although this may seem as a costly sacrifice, it is not, as empirical covers have been argued to be incognizable [6], which prevents model-preserving approaches from being perfectly secure as well.

While the relationship between distortion and steganographic security is far from clear, embedding while minimizing a distortion function is an easier problem than embedding with a steganographic constraint (preserving the distribution of covers). It is also more flexible, allowing the results obtained from experiments with blind steganalyzers to drive the design of the distortion function. In fact, today's least detectable steganographic schemes for digital images [55, 93, 71, 65] were designed using this principle. Moreover, when the distortion is defined as a norm between feature vectors extracted from cover and stego objects, minimizing distortion becomes tightly connected with model preservation insofar the features can be considered as a low-dimensional model of covers. This line of reasoning already appeared in [56, 65] and was further developed in [23].

With the exception of [23], steganographers work with additive distortion functions obtained as a sum of single-letter distortions. A well-known example is matrix embedding where the sender minimizes the total number of embedding changes. Near-optimal coding schemes for this problem appeared in [31, 21], together with other clever constructions and extensions [98, 94, 96, 22, 97, 95]. When the single-letter distortions vary across the cover elements, reflecting thus different costs of individual embedding changes, current coding methods are highly suboptimal [55, 71].

5.1. Distortion function.

For concreteness, and without loss of generality, $\mathbf{x}$ will be called an image and x_i its ith pixel, even though other interpretations are certainly possible. For example, x_i may represent an RGB triple in a color image, a quantized DCT coefficient in a JPEG file, etc. Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X} = \{\mathcal{I}\}^n$ be an n-pixel cover image with the pixel dynamic range $\mathcal{I}$. For example, $\mathcal{I} = \{0, \ldots, 255\}$ for 8-bit grayscale images.

The sender communicates a message to the receiver by introducing modifications to the cover image and sending a stego image $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{Y} = \mathcal{I}_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n$, where $\mathcal{I}_i \subset \mathcal{I}$ are such that $x_i \in \mathcal{I}_i$. The embedding operation is called *binary* if $|\mathcal{I}_i| = 2$, or *ternary* if $|\mathcal{I}_i| = 3$ for every pixel i. For example, ± 1 embedding (sometimes called LSB matching) can be represented by $\mathcal{I}_i = \{x_i - 1, x_i, x_i + 1\}$ with appropriate modifications at the boundary of the dynamic range.

The impact of embedding modifications will be measured using a distortion function D. The sender will strive to embed payload while minimizing D. This section is limited to an additive D in the form[13]

[13] The case of embedding with non-additive distortion functions is addressed in [23] by converting it to a sequence of embeddings with an additive distortion.

$$D(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} \rho_i(\mathbf{x}, y_i), \tag{5.1}$$

where $\rho_i : \mathcal{X} \times \mathcal{I}_i \to [-K, K]$, $0 < K < \infty$, are bounded functions expressing the cost of replacing the cover pixel x_i with y_i. Note that ρ_i may arbitrarily depend on the entire cover image $\mathbf{x}$, allowing thus the sender to incorporate inter-pixel dependencies [65]. The fact that the value of $\rho_i(\mathbf{x}, y_i)$ is independent of changes made at other pixels implies that the embedding changes do not interact.

The boundedness of $D(\mathbf{x},\mathbf{y})$ is not limiting the sender in practice since the case when a particular value y_i is forbidden (a requirement often found in practical steganographic schemes [37]) can be resolved by excluding y_i from $\mathcal{I}_i$. In practice, the sets $\mathcal{I}_i$, $i \in \{1,\dots,n\}$, may depend on cover pixels and thus may not be available to the receiver. To handle this case, the domain of ρ_i is expanded to $\mathcal{X} \times \mathcal{I}$ and defined $\rho_i(\mathbf{x}, y_i) = \infty$ whenever $y_i \notin \mathcal{I}_i$.

The definition of the distortion function is intentionally kept rather general. In particular, it is *not* required that $\rho_i(\mathbf{x}, x_i) \le \rho_i(\mathbf{x}, y_i)$ for all $y_i \in \mathcal{I}_i$ to allow for the case when it is actually beneficial to make an embedding change instead of leaving the pixel unchanged. An example of this situation appears in [23].

5.2. **Problem formulation.** This section contains a formal definition of the problem of embedding while minimizing a distortion function. The performance bounds are stated and some numerical quantities are defined that will be used to compare the coding methods w.r.t. each other and to the bounds.

It is assumed that the sender obtains her payload in the form of a pseudo-random bit stream, such as by compressing or encrypting the original message. Moreover, the embedding algorithm associates every cover image $\mathbf{x}$ with a pair $\{\mathcal{Y}, \pi\}$, where $\mathcal{Y}$ is the set of all stego images into which $\mathbf{x}$ can be modified and π is their probability distribution characterizing the sender's actions, $\pi(\mathbf{y}) \triangleq P(\mathbf{Y} = \mathbf{y}|\mathbf{x})$. Since the choice of $\{\mathcal{Y}, \pi\}$ depends on the cover image, all concepts derived from these quantities necessarily depend on $\mathbf{x}$ as well. Thinking of $\mathbf{x}$ as a constant parameter that is *fixed in the very beginning*, the dependency on $\mathbf{x}$ is not made explicit. For this reason, one can simply write $D(\mathbf{y}) \triangleq D(\mathbf{x},\mathbf{y})$.

If the receiver knew $\mathbf{x}$, the sender could send up to

$$H(\pi) = -\sum_{\mathbf{y}\in\mathcal{Y}} \pi(\mathbf{y}) \log \pi(\mathbf{y}) \tag{5.2}$$

bits on average while introducing the average distortion

$$E_\pi[D] = \sum_{\mathbf{y}\in\mathcal{Y}} \pi(\mathbf{y}) D(\mathbf{y}) \tag{5.3}$$

by choosing the stego image according to π. By the Gel'fand–Pinsker theorem [39], the knowledge of $\mathbf{x}$ does not give any fundamental advantage to the receiver and the same performance can be achieved as long as $\mathbf{x}$ is known to the sender. Indeed, none of the practical embedding algorithms introduced in this report requires the knowledge of $\mathbf{x}$ or D for reading the message.

The task of embedding while minimizing distortion can assume two forms:

- **Payload-limited sender (PLS):** embed a *fixed average payload* of m bits while minimizing the average distortion,

$$\underset{\pi}{\text{minimize}}\ E_\pi[D] \qquad \text{subject to } H(\pi) = m. \tag{5.4}$$

- **Distortion-limited sender (DLS):** maximize the average payload while introducing a *fixed average distortion* D_ϵ,

$$\underset{\pi}{\text{maximize}}\ H(\pi) \qquad \text{subject to } E_\pi[D] = D_\epsilon. \tag{5.5}$$

The problem of embedding a fixed-size message while minimizing the total distortion D (the PLS) is more commonly used in steganography when compared to the DLS. When the distortion function is content-driven, the sender may choose to maximize the payload with a constraint on the overall distortion. This DLS corresponds to a more intuitive use of steganography since images with different level of noise and texture can carry different amount of hidden payload and thus the distortion should be fixed instead of the payload (as long as the distortion corresponds to statistical detectability). The fact that the payload is driven by the image content is essentially a case of the batch-steganography paradigm [50].

5.2.1. *Performance bounds and comparison metrics.* Both embedding problems described above bear relationship to the problem of source coding with a fidelity criterion as described by Shannon [76] and the problem of source coding with side information available at the transmitter, the so-called Gel'fand-Pinsker problem [39]. Problems (5.4) and (5.5) are dual to each other, meaning that the optimal distribution for the first problem is, for some value of D_ϵ, also optimal for the second one. Following the maximum entropy principle [13, Th. 12.1.1], the optimal solution has the form of a Gibbs distribution (see Appendix A in [31] for derivation):

$$\pi(\mathbf{y}) = \frac{\exp(-\lambda D(\mathbf{y}))}{Z(\lambda)} \overset{(a)}{=} \prod_{i=1}^{n} \frac{\exp(-\lambda \rho_i(y_i))}{Z_i(\lambda)} \triangleq \prod_{i=1}^{n} \pi_i(y_i), \tag{5.6}$$

where the parameter $\lambda \in [0, \infty)$ is obtained from the corresponding constraints (5.4) or (5.5) by solving an algebraic equation;[14] $Z(\lambda) = \sum_{\mathbf{y} \in \mathcal{Y}} \exp(-\lambda D(\mathbf{y}))$, $Z_i(\lambda) = \sum_{y_i \in \mathcal{I}_i} \exp(-\lambda \rho_i(y_i))$ are the corresponding partition functions. Step (a) follows from the additivity of D, which also leads to mutual independence of individual stego pixels y_i given $\mathbf{x}$.

By changing each pixel i with probability π_i (5.6) one can *simulate* embedding with optimal π. This is important for steganography developers who can test the security of a scheme that uses the pair $\{\mathcal{Y}, \pi\}$ using blind steganalysis without having to implement a practical embedding algorithm. The simulator of optimal embedding can also be used to assess the increase in statistical detectability of a practical (suboptimal) algorithm w.r.t. to the optimal one. This separation principle simplifies the search for better distortion measures since only the most promising approaches can be implemented. The simulators are used to benchmark different coding algorithms by comparing the security of practical schemes using blind steganalysis.

An established way of evaluating coding algorithms in steganography is to compare the *embedding efficiency* $e(\alpha) = \alpha n / E_\pi[D]$ (in bits per unit distortion) for a fixed expected relative payload $\alpha = m/n$ with the upper bound derived from (5.6). When the number of changes is minimized, e is the average number of bits hidden per embedding change. For general functions ρ_i, the interpretation of this metric becomes less clear. A different and more easily interpretable metric is to compare the payload, m, of an embedding algorithm w.r.t. the payload, m_{MAX}, of the optimal DLS for a fixed D_ϵ,

$$l(D_\epsilon) = \frac{m_{\text{MAX}} - m}{m_{\text{MAX}}}, \tag{5.7}$$

which will be called the *coding loss*.

5.2.2. *Binary embedding operation.* For binary embedding operations, it is enough to consider a slightly narrower class of distortion functions without experiencing any loss of generality. The binary case is important as the embedding method described here can be extended to non-binary operations [20].

For binary embedding with $\mathcal{I}_i = \{x_i, \overline{x}_i\}$, $x_i \neq \overline{x}_i$, let $\rho_i^{\min} = \min\{\rho_i(\mathbf{x}, x_i), \rho_i(\mathbf{x}, \overline{x}_i)\}$ and $\varrho_i = |\rho_i(\mathbf{x}, x_i) - \rho_i(\mathbf{x}, \overline{x}_i)| \geq 0$. Eq. (5.1) can now be rewritten as:

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \rho_i^{\min} + \sum_{i=1}^{n} \varrho_i \cdot [\rho_i^{\min} < \rho_i(\mathbf{x}, y_i)]. \tag{5.8}$$

Because the first sum does not depend on $\mathbf{y}$, when minimizing D over $\mathbf{y}$ it is enough to consider only the second term. It now becomes clear that embedding in cover $\mathbf{x}$ while minimizing (5.8) is equivalent to embedding in cover $\mathbf{z}$

$$z_i = \begin{cases} x_i & \text{when } \rho_i^{\min} = \rho_i(\mathbf{x}, x_i) \\ \overline{x}_i & \text{when } \rho_i^{\min} = \rho_i(\mathbf{x}, \overline{x}_i). \end{cases} \tag{5.9}$$

while minimizing

$$\tilde{D}(\mathbf{z}, \mathbf{y}) = \sum_{i=1}^{n} \tilde{\rho}_i(\mathbf{z}, y_i) \triangleq \sum_{i=1}^{n} \varrho_i \cdot [y_i \neq z_i], \tag{5.10}$$

with non-negative costs $\tilde{\rho}_i(\mathbf{z}, z_i) = 0 \leq \tilde{\rho}_i(\mathbf{z}, \overline{z}_i) = \varrho_i$ for all i (when the cover pixel z_i is changed to $\overline{z}_i$, the distortion $\tilde{D}$ always increases). Thus, from now on binary embedding operations will always consider distortion functions of the form:

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \varrho_i \cdot [y_i \neq x_i], \tag{5.11}$$

with $\varrho_i \geq 0$.

[14] A simple binary search will do the job because both $H(\pi)$ and $E_\pi[D]$ are monotone w.r.t. λ.

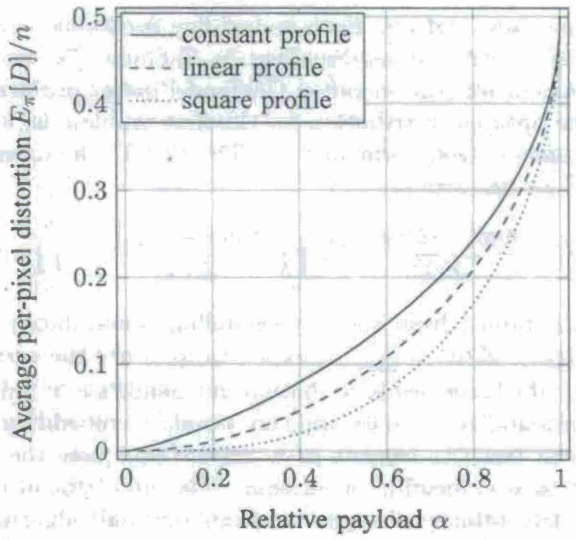

FIGURE 5.1. Lower bound on the average per-pixel distortion, $E_\pi[D]/n$, as a function of relative payload α for different distortion profiles.

For example, F5 [90] uses the distortion function (5.11) with $\varrho_i = 1$ (the number of embedding changes), while nsF5 [37] employs wet paper codes, where $\varrho_i \in \{1, \infty\}$. In some embedding algorithms [33, 55, 71], where the cover is preprocessed and quantized before embedding, ϱ_i is proportional to the quantization error at pixel x_i.

Additionally, for binary embedding operations one introduces the so-called *distortion profile* ϱ if $\varrho_i = \varrho(i/n)$ for all i, where ϱ is a non-decreasing[15] function $\varrho : [0,1] \to [0,K]$. The following distortion profiles are of interest in steganography (this is not an exhaustive list): the *constant profile*, $\varrho(x) = 1$, when all pixels have the same impact on detectability when changed; the *linear profile*, $\varrho(x) = 2x$, when the distortion is related to a quantization error uniformly distributed on $[-Q/2, Q/2]$ for some quantization step $Q > 0$; and the *square profile*, $\varrho(x) = 3x^2$, which can be encountered when the distortion is related to a quantization error that is not uniformly distributed.

The profile ϱ is usually normalized so that $E_\pi[D]/n = \sum_{i=1}^n \pi_i \varrho_i / n = 0.5$ when embedding a full payload $m = n$. With this convention, Figure 5.1 displays the lower bounds on the average per-pixel distortion for three distortion profiles.

In practice, some cover pixels may require $\mathcal{I}_i = \{x_i\}$ and thus $\varrho_i = \infty$ (the so-called *wet pixels* [33, 35, 37]) to prevent the embedding algorithm from modifying them. Since such pixels are essentially constant, in this case one should measure the relative payload α with respect to the set of *dry pixels* $\{x_i | \varrho_i < \infty\}$, i.e., $\alpha = m/|\{x_i|\varrho_i < \infty\}|$. The overall channel is called the wet paper channel and it is characterized by the profile ϱ of dry pixels and *relative wetness* $\tau = |\{x_i|\varrho_i = \infty\}|/n$. The wet paper channel is often required when working with images in the JPEG domain [37].

5.3. **Syndrome coding.** The PLS and the DLS can be realized in practice using a general methodology called *syndrome coding*. This section briefly reviews this approach and its history while Section 5.4 explains the main contribution – the syndrome-trellis codes.

Let us first assume a binary version of both embedding problems. Let $\mathcal{P} : \mathcal{I}_i \to \{0,1\}$ be a parity function shared between the sender and the receiver satisfying $\mathcal{P}(x_i) \neq \mathcal{P}(y_i)$ such as $\mathcal{P}(x) = x \bmod 2$. The sender and the receiver need to implement the embedding and extraction mappings defined as $\mathrm{Emb} : \mathcal{X} \times \{0,1\}^m \to \mathcal{Y}$ and $\mathrm{Ext} : \mathcal{Y} \to \{0,1\}^m$ satisfying

$$\mathrm{Ext}(\mathrm{Emb}(\mathbf{x}, \mathbf{m})) = \mathbf{m} \quad \forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{m} \in \{0,1\}^m,$$

respectively. In particular, the knowledge of the distortion function D at the receiver is not assumed and thus the embedding scheme can be seen as being universal in this sense. A common information-theoretic strategy for solving the PLS problem is known as binning, which is here implemented using cosets of a linear code. Such a construction, better known as syndrome coding, is capacity achieving for the PLS problem if random linear codes are used.

[15] By reindexing the pixels, one can indeed assume that $\varrho_1 \leq \varrho_2 \leq \cdots \leq \varrho_n \leq K$.

In syndrome coding, the embedding and extraction mappings are realized using a binary linear code $\mathcal{C}$ of length n and dimension $n-m$:

$$\text{Emb}(\mathbf{x},\mathbf{m}) = \arg \min_{\mathcal{P}(\mathbf{y})\in\mathcal{C}(\mathbf{m})} D(\mathbf{x},\mathbf{y}), \tag{5.12}$$

$$\text{Ext}(\mathbf{y}) = \mathbb{H}\mathcal{P}(\mathbf{y}), \tag{5.13}$$

where $\mathcal{P}(\mathbf{y}) = (\mathcal{P}(y_1),\dots,\mathcal{P}(y_n))$, $\mathbb{H}\in\{0,1\}^{m\times n}$ is a parity-check matrix of the code $\mathcal{C}$, $\mathcal{C}(\mathbf{m}) = \{\mathbf{z}\in\{0,1\}^n|\mathbb{H}\mathbf{z}=\mathbf{m}\}$ is the coset corresponding to syndrome $\mathbf{m}$, and all operations are in binary arithmetic.

Unfortunately, random linear codes are not practical due to the exponential complexity of the optimal binary coset quantizer (5.12), which is the most challenging part of the problem. STCs form a rich class of codes for which the quantizer can be solved optimally with linear time and space complexity w.r.t. n.

Since the DLS is a dual problem to the PLS, it can be solved by (5.12) and (5.13) once an appropriate message size m is known. This can be obtained in practice by $m = m_{\text{MAX}}(1-l')$, where $m_{\text{MAX}} = H(\pi_\lambda)$ is the maximal average payload obtained from the optimal distribution (5.6) achieving average distortion D_ϵ and l' is an experimentally-obtained coding loss one expects the algorithm will achieve.

5.3.1. *Prior Art.* The problem of minimizing the embedding impact in steganography, introduced above as the PLS problem, has been already conceptually described by Crandall [15] in his essay posted on the steganography mailing list in 1998. He suggested that whenever the encoder embeds at most one bit per pixel, it should make use of the embedding impact defined for every pixel and minimize its total sum:

> "Conceptually, the encoder examines an area of the image and weights each of the options that allow it to embed the desired bits in that area. It scores each option for how conspicuous it is and chooses the option with the best score."

Later, Bierbrauer [3, 4] studied a special case of this problem and described a connection between codes (not necessarily linear) and the problem of minimizing the number of changed pixels (the constant profile). This connection, which has become known as matrix embedding (encoding), was made famous among steganographers by Westfeld [90] who incorporated it in his F5 algorithm. A binary Hamming code was used to implement the syndrome-coding scheme for the constant profile. Later on, different authors suggested other linear codes, such as Golay [83], BCH [75], random codes of small dimension [38], and non-linear codes based on the idea of a blockwise direct sum [4]. Current state-of-the-art methods use codes based on Low Density Generator Matrices (LDGMs) [31] in combination with the ZZW construction [95]. The embedding efficiency of these codes stays rather close to the bound for arbitrarily small relative payloads [28].

The versatile syndrome-coding approach can also be used to communicate via the wet paper channel using the so-called wet paper codes [33]. Wet paper codes minimizing the number of changed dry pixels were described in [34, 75, 97, 22].

Even though other distortion profiles, such as the linear profile, are of great interest to steganography, no general solution with performance close to the bound is currently known. The authors of [55] approached the PLS problem by minimizing the distortion on a block-by-block basis utilizing a Hamming code and a suboptimal quantizer implemented using a brute-force search that allows up to three embedding changes. Such an approach, however, provides highly suboptimal performance far from the theoretical bound (see Figure 5.8). A similar approach based on BCH codes and a brute-force quantizer was described in [71] achieving a slightly better performance than Hamming codes. Neither Hamming or BCH codes can be used to deal with the wet paper channel without significant performance loss. To the best of PI's, no solution is known that could be used to solve the PLS problem with arbitrary distortion profile containing wet pixels.

5.4. **Syndrome-Trellis Codes.** In this section, the focus is on solving the binary PLS problem with distortion function (5.10). The resulting codes are called the syndrome-trellis codes. These codes will serve as a building block for non-binary PLS and DLS problems as described in [20].

The construction behind STCs is not new from an information-theoretic perspective, since the STCs are convolutional codes represented in a dual domain. However, STCs are very interesting for practical steganography since they allow solving both embedding problems with a very small coding loss over a wide range of distortion profiles even with wet pixels. The same code can be used with all profiles making the embedding algorithm practically universal. STCs offer general and state-of-the-art solution for both embedding problems in steganography. Here, the PI gives the description of the codes along with their graphical representation, the syndrome trellis. Such construction is prepared for the Viterbi algorithm, which is optimal for solving (5.12). Important practical guidelines for optimizing the codes and using them for the wet paper channel are also covered. Finally, the PI studies the performance of these codes by extensive numerical simulations using different distortion profiles including the wet paper channel.

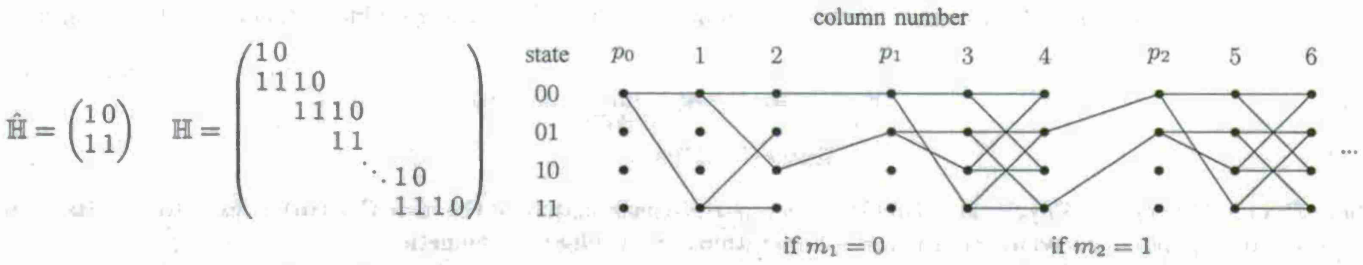

FIGURE 5.2. Example of a parity-check matrix $\mathbb{H}$ formed from the submatrix $\hat{\mathbb{H}}$ ($h = 2, w = 2$) and its corresponding syndrome trellis. The last $h-1$ submatrices in $\mathbb{H}$ are cropped to achieve the desired relative payload α. The syndrome trellis consists of repeating blocks of $w+1$ columns, where "p_0" and "p_i", $i > 0$, denote the starting and pruning columns, respectively. The column labeled $l \in \{1, 2, \ldots\}$ corresponds to the lth column in the parity-check matrix $\mathbb{H}$.

The main goal is to develop efficient syndrome-coding schemes for an *arbitrary* relative payload α with the main focus on small relative payloads (think of $\alpha \leq 1/2$ for example). In steganography, the relative payload must decrease with increasing size of the cover object in order to maintain the same level of security, which is a consequence of the square root law [27]. Moreover, recent results from steganalysis in both spatial [64] and DCT domains [57] suggest that the secure payload for digital image steganography is always far below 1/2. Another reason for targeting smaller payloads is the fact that as $\alpha \to 1$, all binary embedding algorithms tend to introduce changes with probability 1/2, no matter how optimal they are. Denoting with $R = (n-m)/n$ the rate of the linear code $\mathcal{C}$, then $\alpha \to 0$ translates to $R = 1 - \alpha \to 1$, which is characteristic for applications of syndrome coding in steganography.

5.4.1. *From convolutional codes to syndrome-trellis codes.* Since Shannon [76] introduced the problem of source coding with a fidelity criterion in 1959, convolutional codes were probably the first "practical" codes used for this problem [86]. This is because the gap between the bound on the expected per-pixel distortion and the distortion obtained using the optimal encoding algorithm (the Viterbi algorithm) decreases exponentially with the constraint length of the code [86, 44]. The complexity of the Viterbi algorithm is linear in the block length of the code, but exponential in its constraint length (the number of trellis states grows exponentially in the constraint length).

When adapted to the PLS problem, convolutional codes can be used for syndrome coding since the best stego image in (5.12) can be found using the Viterbi algorithm. This makes convolutional codes (of small constraint length) suitable for steganography because the entire cover object can be used and the speed can be traded for performance by adjusting the constraint length. Note that the receiver does not need to know D since only the Viterbi algorithm requires this knowledge. By increasing the constraint length, one can achieve the average per-pixel distortion that is arbitrarily close to the bounds and thus make the coding loss (5.7) approach zero. Convolutional codes are often represented with shift-registers (see Chapter 48 in [59]) that generate the codeword from a set of information bits. In channel coding, codes of rates $R = 1/k$ for $k = 2, 3, \ldots$ are usually considered for their simple implementation.

Convolutional codes in standard trellis representation are commonly used in problems that are dual to the PLS problem, such as the distributed source coding. The main drawback of convolutional codes, when implemented using shift-registers, comes from our requirement of small relative payloads (code rates close to one) which is specific to steganography. A convolutional code of rate $R = (k-1)/k$ requires $k-1$ shift registers in order to implement a scheme for $\alpha = 1/k$. Here, unfortunately, the complexity of the Viterbi algorithm in this construction grows exponentially with k. Instead of using puncturing (see Chapter 48 in [59]), which is often used to construct high-rate convolutional codes, the PI prefers to represent the convolutional code in the dual domain using its parity-check matrix. In fact, Sidorenko and Zyablov [78] showed that optimal decoding of convolutional codes (our binary quantizer) with rates $R = (k-1)/k$ can be carried out in the dual domain on the syndrome trellis with a much lower complexity and without any loss of performance. This approach is more efficient as $\alpha \to 0$ and thus is choosen for the construction.

In the dual domain, a code of length n is represented by a parity-check matrix instead of a generator matrix as is more common for convolutional codes. Working directly in the dual domain allows the Viterbi algorithm to exactly implement the coset quantizer required for the embedding function (5.12). The message can be extracted in a straightforward manner by the recipient using the shared parity-check matrix.

5.5. **Description of syndrome-trellis codes.** Although syndrome-trellis codes form a class of convolutional codes and thus can be described using a classical approach with shift-registers, it is advantageous to stay in the dual domain and

Forward part of the Viterbi algorithm

```
wght[0] = 0
wght[1,...,2^h-1] = infinity
indx = indm = 1
for i = 1,...,num of blocks (submatrices in H) {
 for j = 1,...,w {                  // for each column
  for k = 0,...,2^h-1 {             // for each state
   w0 = wght[k] + x[indx]*rho[indx]
   w1 = wght[k XOR H_hat[j]] + (1-x[indx])*rho[indx]
   path[indx][k] = w1 < w0 ? 1 : 0      // C notation
   newwght[k] = min(w0, w1)
  }
  indx++
  wght = newwght
 }
  // prune states
 for j = 0,...,2^(h-1)-1
  wght[j] = wght[2*j + message[indm]]
 wght[2^(h-1),...,2^h-1] = infinity
 indm++
}
```

Backward part of the Viterbi alg.

```
embedding_cost = wght[0]
state = 0, indx--, indm--
for i = num of blocks,...,1 (step -1) {
  for j = w,...,1 (step -1) {
    y[indx] = path[indx][state]
    state = state XOR (y[indx]*H_hat[j])
    indx--
  }
  state = 2*state + message[indm]
  indm--
}
```

Legend

```
INPUT: x, message, H_hat
  x = (x[1],...,x[n]) cover object
  message = (message[1],...,message[m])
  H_hat[j] = j th column in int notation

OUTPUT: y, embedding_cost
  y = (y[1],...,y[n]) stego object
```

FIGURE 5.3. Pseudocode of the Viterbi algorithm modified for the syndrome trellis.

describe the code directly by its parity-check matrix. The parity-check matrix $\mathbb{H} \in \{0,1\}^{m \times n}$ of a binary syndrome-trellis code of length n and codimension m is obtained by placing a small submatrix $\hat{\mathbb{H}}$ of size $h \times w$ along the main diagonal as in Figure 5.2. The submatrices $\hat{\mathbb{H}}$ are placed next to each other and shifted down by one row leading to a sparse and banded $\mathbb{H}$. The height h of the submatrix (called the *constraint height*) is a design parameter that affects the algorithm speed and efficiency (typically, $6 \leq h \leq 15$). The width of $\hat{\mathbb{H}}$ is dictated by the desired ratio of m/n, which coincides with the relative payload $\alpha = m/n$ when no wet pixels are present. If m/n equals to $1/k$ for some $k \in \mathbb{N}$, select $w = k$. For general ratios, find k such that $1/(k+1) < m/n < 1/k$. The matrix $\mathbb{H}$ will contain a mix of submatrices of width k and $k+1$ so that the final matrix $\mathbb{H}$ is of size $m \times n$. In this way, one can create a parity-check matrix for an arbitrary message and code size. The submatrix $\hat{\mathbb{H}}$ acts as an input parameter shared between the sender and the receiver and its choice is discussed in more detail in Section 5.5.2. For the sake of simplicity, in the following description it is assumed that $m/n = 1/w$ and thus the matrix $\mathbb{H}$ is of the size $b \times (b \cdot w)$, where b is the number of copies of $\hat{\mathbb{H}}$ in $\mathbb{H}$.

Similar to convolutional codes and their trellis representation, every codeword of an STC $\mathcal{C} = \{\mathbf{z} \in \{0,1\}^n | \mathbb{H}\mathbf{z} = 0\}$ can be represented as a unique path through a graph called the *syndrome trellis*. Moreover, the syndrome trellis is parametrized by $\mathbf{m}$ and thus can represent members of arbitrary coset $\mathcal{C}(\mathbf{m}) = \{\mathbf{z} \in \{0,1\}^n | \mathbb{H}\mathbf{z} = \mathbf{m}\}$. An example of the syndrome trellis is shown in Figure 5.2. More formally, the syndrome trellis is a graph consisting of b blocks, each containing $2^h(w+1)$ nodes organized in a grid of $w+1$ columns and 2^h rows. The nodes between two adjacent columns form a bipartite graph, i.e., all edges only connect nodes from two adjacent columns. Each block of the trellis represents one submatrix $\hat{\mathbb{H}}$ used to obtain the parity-check matrix $\mathbb{H}$. The nodes in every column are called *states*.

Each $\mathbf{z} \in \{0,1\}^n$ satisfying $\mathbb{H}\mathbf{z} = \mathbf{m}$ is represented as a path through the syndrome trellis which represents the process of calculating the syndrome as a linear combination of the columns of $\mathbb{H}$ with weights given by $\mathbf{z}$. Each path starts in the leftmost all-zero state in the trellis and extends to the right. The path shows the step-by-step calculation of the (partial) syndrome using more and more bits of $\mathbf{z}$. For example, the first two edges in Figure 5.2, that connect the state 00 from column p_0 with states 11 and 00 in the next column, correspond to adding ($\mathcal{P}(y_1) = 1$) or not adding ($\mathcal{P}(y_1) = 0$) the first column of $\mathbb{H}$ to the syndrome, respectively.[16] At the end of the first block, all paths for which the first bit of the partial syndrome does not match m_1 are terminated. This way, one obtains a new column of the trellis, which will serve as the starting column of the next block. This column merely illustrates the transition of the trellis from representing the partial syndrome $(s_1, \ldots, s_h)$ to $(s_2, \ldots, s_{h+1})$. This operation is repeated at each block transition in the matrix $\mathbb{H}$ and guarantees that 2^h states are sufficient to represent the calculation of the partial syndrome throughout the whole syndrome trellis.

[16]The state corresponds to the partial syndrome.

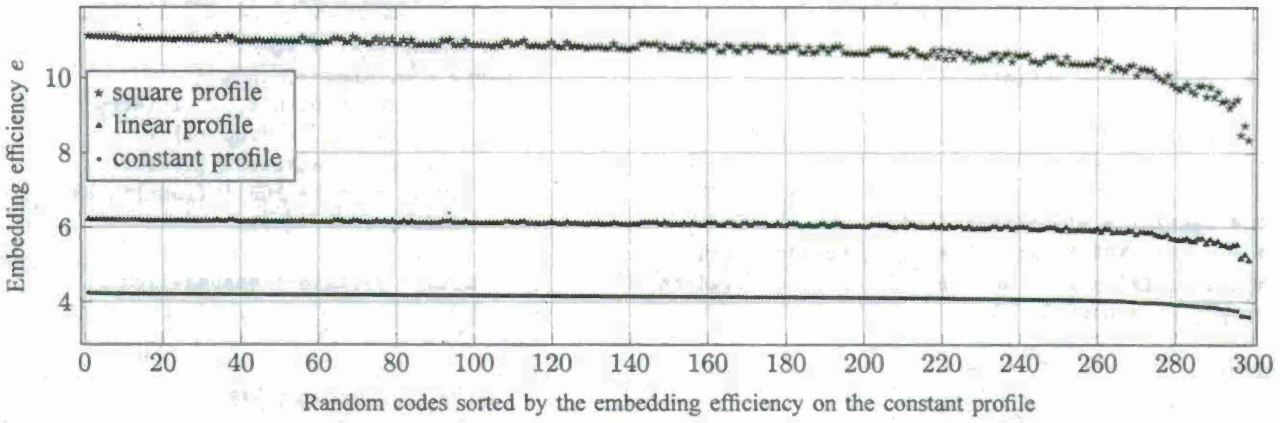

FIGURE 5.4. Embedding efficiency of 300 random syndrome-trellis codes satisfying the design rules for relative payload $\alpha = 1/2$ and constraint height $h = 10$. All codes were evaluated by the Viterbi algorithm with a random cover object of $n = 10^6$ pixels and a random message on the constant, linear, and square profiles. Codes are shown in the order determined by their embedding efficiency evaluated on the constant profile. This experiment suggests that codes good for the constant profile are good for other profiles. Codes designed for different relative payloads have a similar behavior.

To find the closest stego object, weights are assigned to all trellis edges. The weights of the edges entering the column with label l, $l \in \{1, \dots, n\}$, in the syndrome trellis depend on the lth bit representation of the original cover object $\mathbf{x}$, $\mathcal{P}(x_l)$. If $\mathcal{P}(x_l) = 0$, then the horizontal edges (corresponding to not adding the lth column of $\mathbb{H}$) have a weight of 0 and the edges corresponding to adding the lth column of $\mathbb{H}$ have a weight of ϱ_l. If $\mathcal{P}(x_l) = 1$, the roles of the edges are reversed. Finally, all edges connecting the individual blocks of the trellis have zero weight.

The embedding problem (5.12) for binary embedding can now be optimally solved by the *Viterbi algorithm* with time and space complexity $\mathcal{O}(2^h n)$. This algorithm consists of two parts, the *forward* and the *backward* part. The forward part of the algorithm consists of $n + b$ steps. Upon finishing the ith step, one knows the shortest path between the leftmost all-zero state and every state in the ith column of the trellis. Thus in the final, $n + b$th step, one discovers the shortest path through the entire trellis. During the backward part, the shortest path is traced back and the parities of the closest stego object $\mathcal{P}(\mathbf{y})$ are recovered from the edge labels. The Viterbi algorithm modified for the syndrome trellis is described in Figure 5.3 using a pseudocode.

5.5.1. *Implementation details.* The construction of STCs is not constrained to having to repeat the same submatrix $\hat{\mathbb{H}}$ along the diagonal. Any parity-check matrix $\mathbb{H}$ containing at most h nonzero entries along the main diagonal will have an efficient representation by its syndrome trellis and the Viterbi algorithm will have the same complexity $\mathcal{O}(2^h n)$. In practice, the trellis is built on the fly because only the structure of the submatrix $\hat{\mathbb{H}}$ is needed (see the pseudocode in Figure 5.3). As can be seen from the last two columns of the trellis in Figure 5.2, the connectivity between trellis columns is highly regular which can be used to speed up the implementation by "vectorizing" the calculations.

In the forward part of the algorithm, one needs to store one bit (the label of the incoming edge) to be able to reconstruct the path in the backward run. This space complexity is linear and should not cause any difficulty, since for $h = 10$, $n = 10^6$, the total of $2^{10} \cdot 10^6/8$ bytes ($\approx$ 122MB) of space is required. If less space is available, one can always run the algorithm on smaller blocks, say $n = 10^4$, without any noticeable performance drop. If one is only interested in the total distortion $D(\mathbf{y})$ and not the stego object itself, this information does not need to be stored at all and only the forward run of the Viterbi algorithm is required.

5.5.2. *Design of good syndrome-trellis codes.* A natural question regarding practical applications of syndrome-trellis codes is how to optimize the structure of $\hat{\mathbb{H}}$ for fixed parameters h and w and a given profile. If $\hat{\mathbb{H}}$ depended on the distortion profile, the profile would have to be somehow communicated to the receiver. Fortunately, this is not the case and a submatrix $\hat{\mathbb{H}}$ optimized for one profile seems to be good for other profiles as well. In this section, the PI studies these issues experimentally and describes a practical algorithm for obtaining good submatrices.

Let us suppose that the goal is to design a submatrix $\hat{\mathbb{H}}$ of size $h \times w$ for a given constraint height h and relative payload $\alpha = 1/w$. The PI was unable to find a better algorithm than an exhaustive search guided by some simple design rules. First, $\hat{\mathbb{H}}$ should not have identical columns because the syndrome trellis would contain two or more different paths

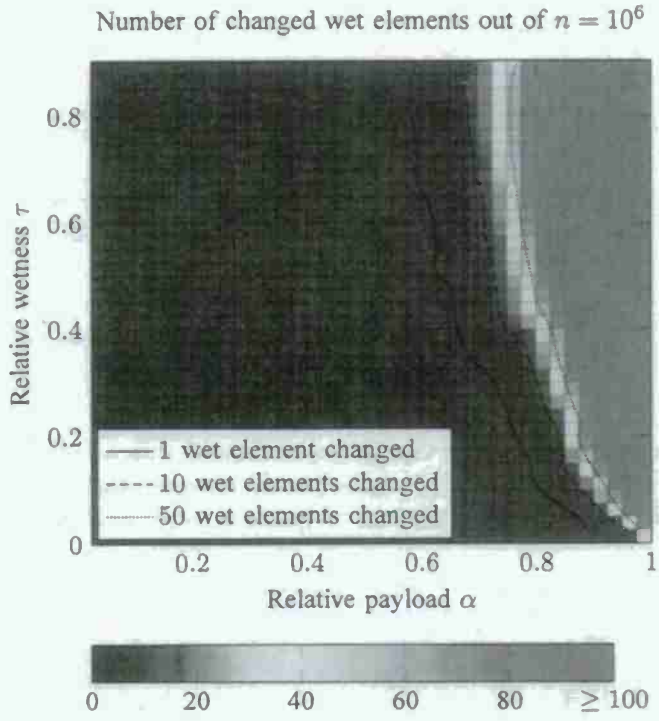

FIGURE 5.5. Average number of wet pixels out of $n = 10^6$ that need to be changed to find a solution to (5.12) using STCs with $h = 11$.

with exactly the same weight, which would lead to an overall decrease in performance. By running an exhaustive search over small matrices, the PI observed that the best submatrices $\hat{\mathbb{H}}$ had ones in the first and last rows. For example, when $h = 7$ and $w = 4$, more than 97% of the best 1000 codes obtained from the exhaustive search satisfied this rule. Thus, the search for good matrices was limited to those that did not contain identical columns and with all bits in the first and last rows set to 1 (the remaining bits were assigned at random). In practice, one can randomly generate $10 - 1000$ submatrices satisfying these rules and estimate their performance (embedding efficiency) experimentally by running the Viterbi algorithm with random covers and messages. For a reliable estimate, cover objects of size at least $n = 10^6$ are required.

To investigate the stability of the design w.r.t. to the profile, the following experiment was conducted. The PI fixed $h = 10$ and $w = 2$, which correspond to a code with $\alpha = 1/2$. The code design procedure was simulated by randomly generating 300 submatrices $\hat{\mathbb{H}}_1, \ldots, \hat{\mathbb{H}}_{300}$ satisfying the above design rules. The goodness of the code was evaluated using the embedding efficiency ($e = m/D(\mathbf{x}, \mathbf{y})$) by running the Viterbi algorithm on a random cover object (of size $n = 10^6$) and with a random message. This was repeated independently for all three profiles from Section 5.2.2. Figure 5.4 shows the embedding efficiency after ordering all 300 codes by their performance on the constant profile. Because the codes with a high embedding efficiency on the constant profile exhibit high efficiency for the other profiles, the code design appears stable w.r.t. the profile and these matrices can be used with other profiles in practice. All further results are generated by using these matrices.

5.5.3. *Wet paper channel.* In this section, the PI investigates how STCs can be used for the wet paper channel described by relative wetness $\tau = |\{i|\rho_i = \infty\}|/n$ with a given distortion profile of dry pixels. Although the STCs can be directly applied to this problem, the probability of not being able to embed a message without changing any wet pixel may be positive and depends on the number of wet pixels, the payload, and the code. The goal is to make this probability very small or to make sure that the number of wet pixels that must be changed is small (e.g., one or two). Two different approaches are now described to address this problem.

Let us assume that the wet channel is iid with probability of a pixel being wet $0 \leq \tau < 1$. This assumption is plausible because the cover pixels can be permuted using a stego key before embedding. For the wet paper channel, the relative payload is defined w.r.t. the dry pixels as $\alpha = m/|\{i|\rho_i < \infty\}|$. When designing the code for the wet paper channel

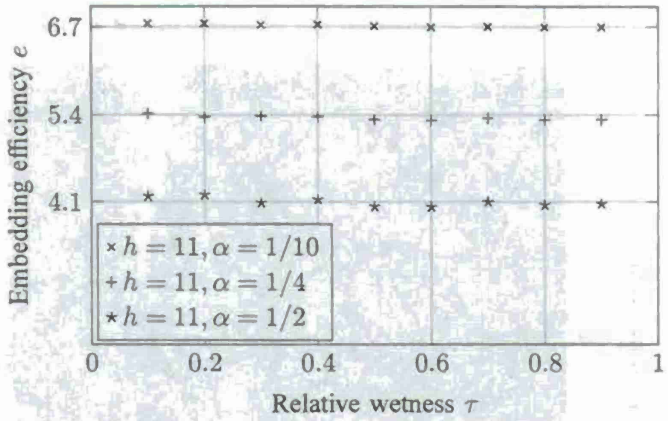

FIGURE 5.6. Effect of relative wetness τ of the wet paper channel with a constant profile on the embedding efficiency of STCs. The distortion was calculated w.r.t. the changed dry pixels only and $\alpha = m/(n - \tau n)$. Each point was obtained by quantizing a random vector of $n = 10^6$ pixels.

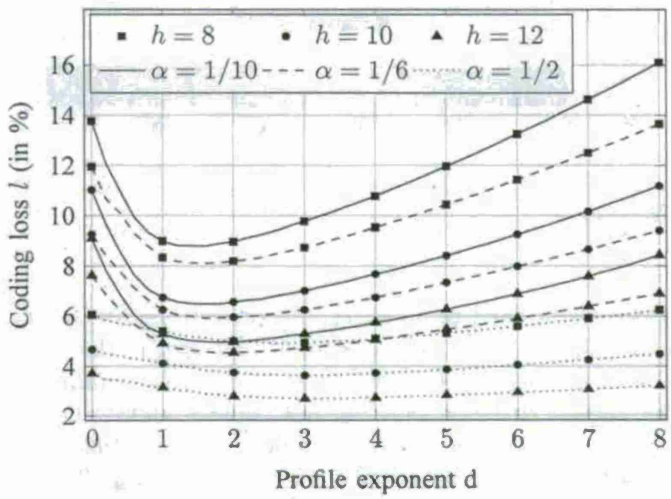

FIGURE 5.7. Comparison of the coding loss of STCs as a function of the profile exponent d for different payloads and constraint heights of STCs. Each point was obtained by quantizing a random vector of $n = 10^6$ pixels.

with n-pixel covers, relative wetness τ, and desired relative payload α, the parity-check matrix $\mathbb{H}$ has to be of the size $[(1-\tau)\alpha n] \times n$.

The random permutation makes the Viterbi algorithm less likely to fail to embed a message without having to change some wet pixels. The probability of failure, p_w, decreases with decreasing α and τ and it also depends on the constraint height h. From practical experiments with $n = 10^6$ cover pixels, $\tau = 0.8$, and $h = 10$, the PI estimated from 1000 independent runs $p_w \doteq 0.24$ for $\alpha = 1/2$, $p_w \doteq 0.009$ for $\alpha = 1/4$, and $p_w \doteq 0$ for $\alpha = 1/10$. In practice, the message size m can be used as a seed for the pseudo-random number generator. If the embedding process fails, embedding $m - 1$ bits leads to a different permutation while embedding roughly the same amount of message. In k trials, the probability of having to modify a wet pixel is at most p_w^k, which can be made arbitrarily small.

Alternatively, the sender may allow a small number of wet pixels to be modified, say one or two, without affecting the statistical detectability in any significant manner. Making use of this fact, one can set the distortion of all wet cover pixels to $\hat{\varrho}_i = C$, $C > \sum_{\varrho_i < \infty} \varrho_i$ and $\hat{\varrho}_i = \varrho_i$ for i dry. The weight c of the best path through the syndrome trellis obtained by the Viterbi algorithm with distortion $\hat{\varrho}_i$ can be written in the form $c = n_c C + c'$, where n_c is the smallest number of wet cover pixels that had to be changed and c' is the smallest weight of the path over the pixels that are allowed to be changed.

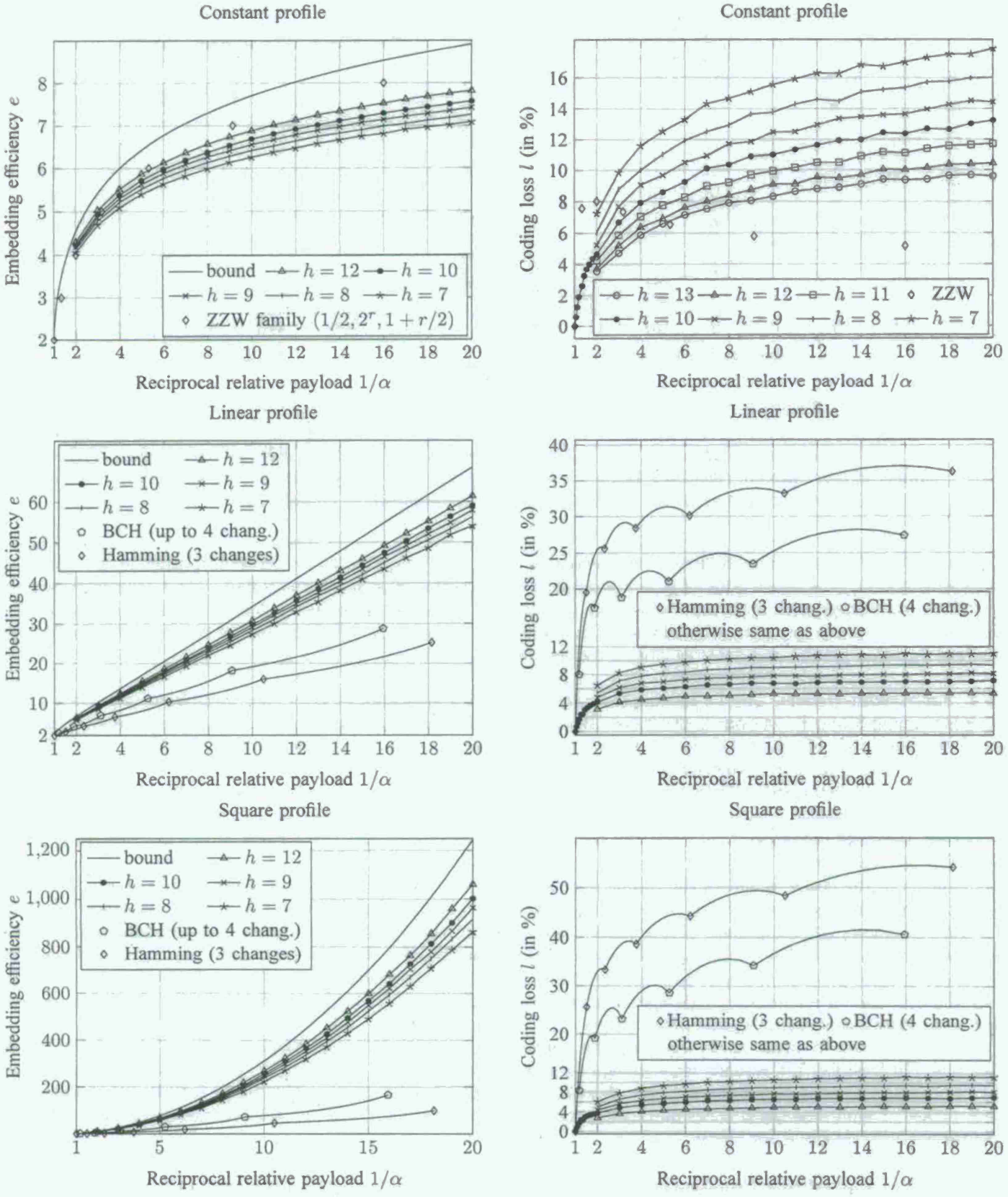

FIGURE 5.8. Embedding efficiency and coding loss of syndrome-trellis codes for three distortion profiles. Each point was obtained by running the Viterbi algorithm with $n = 10^6$ cover pixels. Hamming [55] and BCH [93] codes were applied on a block-by-block basis on cover objects with $n = 10^5$ pixels with a brute-force search making up to three and four changes, respectively. The line connecting a pair of Hamming or BCH codes represents the codes obtained by their block direct sum. For clarity, the PI presents the coding loss results in the range $\alpha \in [0.5, 1]$ only for constraint height $h = 10$ of the syndrome-trellis codes.

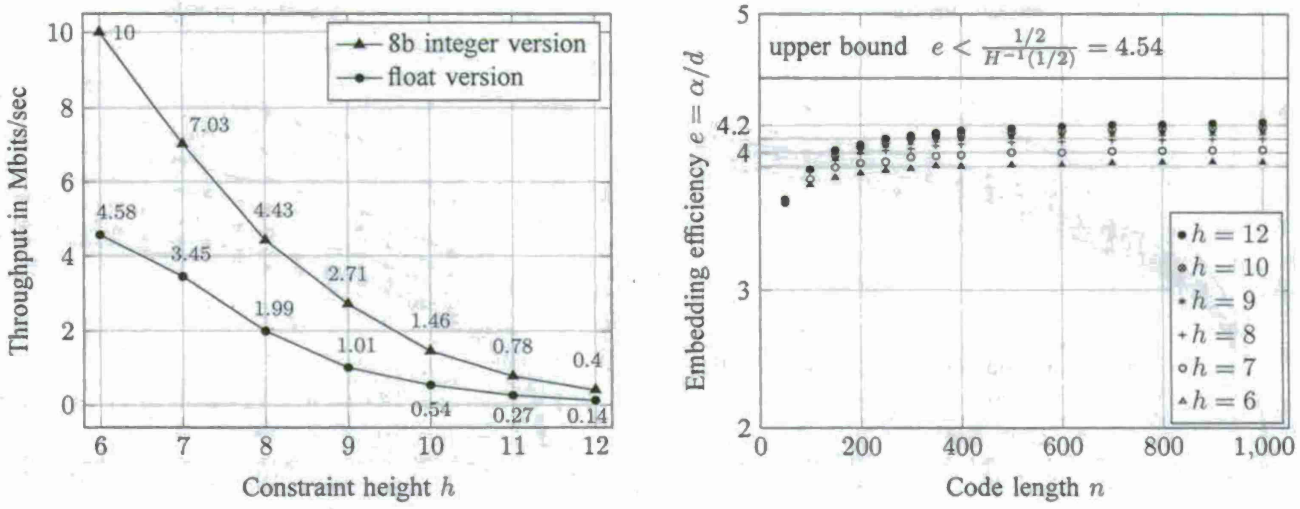

FIGURE 5.9. Results for the syndrome-trellis codes designed for relative payload $\alpha = 1/2$. Left: Average number of cover pixels ($\times 10^6$) quantized per second (throughput) shown for different constraint heights and two different implementations. Right: Average embedding efficiency for different code lengths n (the number of cover pixels), constraint heights h, and a constant distortion profile. Codes of length $n > 1000$ have similar performance as for $n = 1000$. Each point was obtained as an average over 1000 samples.

Figure 5.5 shows the average number of wet pixels out of $n = 10^6$ required to be changed in order to solve (5.12) for STCs with $h = 11$. The exact value of ϱ_i is irrelevant in this experiment as long as it is finite. This experiment suggests that STCs can be used with arbitrary τ as long as $\alpha \leq 0.7$. As can be seen from Figure 5.6, increasing the amount of wet pixels does not lead to any noticeable difference in embedding efficiency for constant profile. Similar behavior has been observed for other profiles and holds as long as the number of changed wet pixels is small.

5.5.4. *Experimental results.* The PI has implemented the Viterbi algorithm in C++ and optimized its performance by using Streaming SIMD Extensions instructions. Based on the distortion profile, the algorithm chooses between the float and 1 byte unsigned integer data type to represent the weight of the paths in the trellis. The following results were obtained using an Intel Core2 X6800 2.93GHz CPU machine utilizing a single CPU core.

Using the search described in Section 5.5.2, the PI found good syndrome-trellis codes of constraint height $h \in \{6, \dots, 12\}$ for relative payloads $\alpha = 1/w$, $w \in \{1, \dots, 20\}$. Some of these codes can be found in [26, Table 1]. In practice, almost every code satisfying the design rules is equally good. This fact can also be seen from Figure 5.4, where 300 random codes are evaluated over different profiles.

The effect of the profile shape on the coding loss for $\varrho(x) \approx x^d$ as a function of d is shown in Figure 5.7. The coding loss increases with decreasing relative payload α. This effect can be compensated by using a larger constraint height h.

Figure 5.8 shows the comparison of syndrome-trellis codes for three profiles with other codes which are known for a given profile. The ZZW family [96] applies only to the constant profile. For a given relative payload α and constraint height h, the same submatrix $\hat{\mathbb{H}}$ was used for all profiles. This demonstrates the versatility of the proposed construction, since the information about the profile does not need to be shared, or, perhaps more importantly, the profile does not need to be known a priori for a good performance.

Figure 5.9 shows the average throughput (the number of cover pixels n quantized per second) based on the used data type. In practice, 1–5 seconds were enough to process a cover object with $n = 10^6$ pixels. The same figure shows the embedding efficiency obtained from very short codes for the constant profile. This result indicates that the average performance of syndrome-trellis codes quickly approaches its maximum w.r.t. n. This is again an advantage, since some applications may require short blocks.

5.6. **Practical Embedding Constructions.** In this section, the PI shows some applications of the proposed methodology for spatial and transform domain (JPEG) steganography. In the past, most embedding schemes were constrained by practical ways of how to encode the message so that the receiver can read it. Problems such as "shrinkage" in F5 [90, 37] or in MMx [55] arose from this practical constraint. By being able to solve the PLS and DLS problems close to the bound

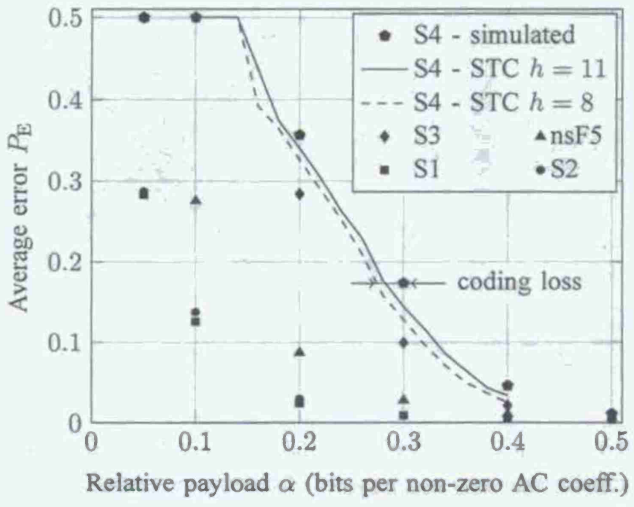

FIGURE 5.10. Comparison of methods with four different weight-assignment strategies S1–S4 and nsF5 as described in Section 5.6.1 when simulated as if the best coding scheme was available. The performance of strategy S4 when practically implemented using STCs with $h = 8$ and $h = 11$ is also shown.

for an arbitrary additive distortion function,[17] steganographers now have much more freedom in designing new embedding algorithms. They only need to select the distortion function and then apply the proposed framework. The only task left to the steganographer is the choice of the distortion function D. It should be selected so that it correlates with statistical detectability. Instead of delving into the difficult problem of how to select the best D, the PI now provides a few examples of additive distortion measures motivated by recent developments in steganography and shows their performance when blind steganalysis is used.

In the examples below, the PI tested the embedding schemes using blind feature-based steganalysis on a large database of images. The image database was evenly divided into a training and a testing set of cover and stego images, respectively. A soft-margin support-vector machine was trained using the Gaussian kernel. The kernel width and the penalty parameter were determined using five-fold cross validation on the grid $(C, \gamma) \in \{(10^k, 2^{j-d}) | k \in \{-3, \dots, 4\}, j \in \{-3, \dots, 3\}\}$, where d is the binary logarithm of the number of features. The results are reported using a measure frequently used in steganalysis – the minimum average classification error

$$P_E = \min_{P_{FA}} (P_{FA} + P_{MD}(P_{FA}))/2, \tag{5.14}$$

where P_{FA} and P_{MD} are the false-alarm and missed-detection probabilities.

5.6.1. *DCT domain steganography.* To apply the proposed framework, one first needs to design an additive distortion function that can be tested by simulating the embedding as if the best codes are available. Finally, the the most promising approach is implemented using STCs. It is assumed that the cover is a grayscale bitmap image which is JPEG compressed to obtain the cover image. Let $\mathcal{A}$ be a set of indices corresponding to AC DCT coefficients after the block-DCT transform and let c_i be the ith AC coefficient before it is quantized with the quantization step q_i for $i \in \mathcal{A}$. Let $\mathcal{X}$ represent the set of all vectors containing quantized AC DCT coefficients divided by their corresponding quantization step. In ordinary JPEG compression, the values c_i are quantized to $x_i \triangleq [c_i/q_i]$.

[Proposed distortion functions]

A binary embedding operation is defined as $\mathcal{I}_i \triangleq \{x_i, \overline{x}_i\}$ by $\overline{x}_i = x_i + \text{sign}(c_i/q_i - x_i)$, where $\text{sign}(x)$ is 1 if $x > 0$, -1 if $x < 0$ and $\text{sign}(0) \in \{-1, 1\}$ uniformly at random. In simple words, x_i is a quantized AC DCT coefficient and $\overline{x}_i$ is the same coefficient when quantized in the opposite direction. Let $e_i = |c_i/q_i - x_i|$ be the quantization error introduced by JPEG compression. By replacing x_i with $\overline{x}_i$ the error becomes $|c_i/q_i - \overline{x}_i| = 1 - e_i$. If $e_i = 0.5$, then the direction where c_i/q_i is rounded depends on the implementation of the JPEG compressor and only small perturbation of the original image may lead to different results. Let $\mathcal{P}(x) = x \bmod 2$. By construction, $\mathcal{P}$ satisfies the property of a parity function, $\mathcal{P}(x_i) \neq \mathcal{P}(\overline{x}_i)$. The distortion function is assumed to be in the form $D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \varrho_i \cdot [x_i \neq y_i]$, where $n = |\mathcal{A}|$.

The following four approaches utilizing values of e_i and q_i were considered. All methods assign $\varrho_i = \infty$ when $c_i/q_i \in (-0.5, 0.5)$ and differ in the definition of the remaining values ϱ_i as follows:

[17] The additivity constraint can be relaxed and more general distortion measures can be used with the PLS and DLS problems in practice [23].

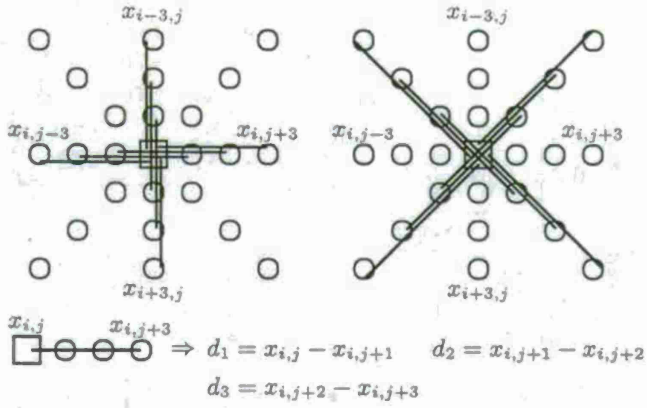

FIGURE 5.11. Set of 4-pixel cliques used for calculating the distortion for digital images represented in the spatial-domain. The final distortion $\rho_{i,j}(y_{i,j})$ is obtained as a sum of terms penalizing the change in pixel $x_{i,j}$ measured w.r.t. each clique containing $x_{i,j}$.

- **S1:** $\varrho_i = 1 - 2e_i$ if $c_i/q_i \notin (-0.5, 0.5)$ (as in perturbed quantization [33]),
- **S2:** $\varrho_i = q_i(1 - 2e_i)$ if $c_i/q_i \notin (-0.5, 0.5)$ (the same as S1 but ϱ_i is weighted by the quantization step),
- **S3:** $\varrho_i = 1$ if $c_i/q_i \in (-1, -0.5] \cup [0.5, 1)$ and $\varrho_i = 1 - 2e_i$ otherwise, and
- **S4:** $\varrho_i = q_i$ if $c_i/q_i \in (-1, -0.5] \cup [0.5, 1)$ and $\varrho_i = q_i(1 - 2e_i)$ otherwise which is similar weight assignment as proposed in [71].

To see the importance of the side-information in the form of the uncompressed cover image, the PI also includes in the tests the nsF5 [37] algorithm, which can be represented using the above formalism as $x_i = [c_i/q_i]$, $\overline{x}_i = x_i - \text{sign}(x_i)$, and $\varrho_i = \infty$ if $x_i = 0$ and $\varrho_i = 1$ otherwise. This way, one always has $|\overline{x}_i| < |x_i|$. The nsF5 embedding minimizes the number of changes to non-zero AC DCT coefficients.

[Steganalysis setup and experimental results]

The proposed strategies were tested on a database of 6, 500 digital camera images prepared as described in [58, Sec. 4.1] so that their smaller size was 512 pixels. The JPEG quality factor 75 was used for compression. The steganalyzer employed the 548-dimensional CC-PEV feature set [57]. Figure 5.10 shows the minimum average classification error P_{E} achieved by simulating each strategy on the bound using the PLS formulation. The strategies S1 and S2, which assign zero cost to coefficients $c_i/q_i = 0.5$, were worse than the nsF5 algorithm that does not use any side-information. On the other hand, strategy S4, which also utilizes the knowledge about the quantization step, was the best. By implementing this strategy, it is necessary to deal with a wet paper channel which can be well modeled by a linear profile with relative wetness $\tau \approx 0.6$ depending on the image content. The PI implemented strategy S4 using STCs, where wet pixels were handled by setting $\varrho_i = C$ for a sufficiently large C. As seen from the results using STCs, payloads below 0.15 bits per non-zero AC DCT coefficient were undetectable using the employed steganalyzer.

Note that the strategy utilized only the information obtainable from a single AC DCT coefficient. In reality, ϱ_i will likely depend on the local image content, quantization errors, and quantization steps. The PI postpones the problem of optimizing D w.r.t. statistical detectability to future research.

5.6.2. *Spatial domain steganography.* To demonstrate the merit of the STC-based multi-layered construction, the PI presents a practical embedding scheme that was largely motivated by [65] and [23]. Single per-pixel distortion function $\rho_{i,j}(y_{i,j})$ should assign the cost of changing the i,jth pixel $x_{i,j}$, first, from its neighborhood and then also based on the new value $y_{i,j}$. Changes made in smooth regions often tend to be highly detectable by blind steganalysis which should lead to high distortion values. On the other hand, pixels which are in busy and hard-to-model regions can be changed more often.

[Proposed distortion functions]

The distortion function is designed based on a model built from a set of all straight 4-pixel lines in 4 different orientations containing the i,jth pixel called cliques (see Figure 5.11). Based on the set of all such cliques, one decides on the value $\rho_{i,j}(y_{i,j})$. Due to strong inter-pixel dependencies, most cliques contain very similar values and thus differences between neighboring pixels tend to be very close to zero. It has been experimentaly observed [65], that the number of cliques falls of quickly as the differences get larger. From this point of view, any clique with small differences should lead to a larger distortion because there are more samples the warden can use for training her steganalyzer.

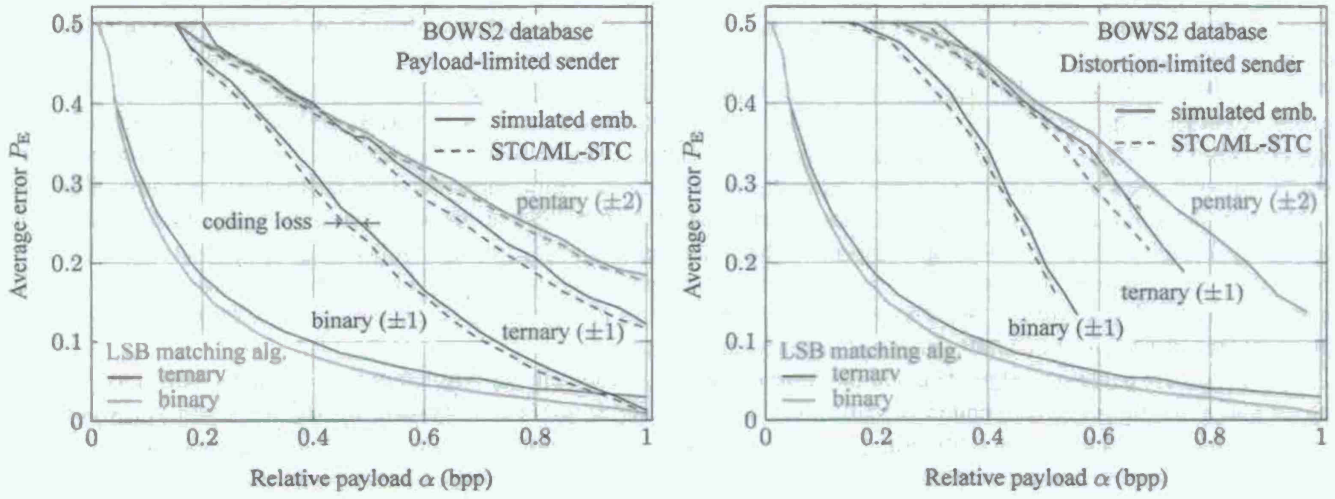

FIGURE 5.12. Comparison of LSB matching with optimal binary and ternary coding with embedding algorithms based on the additive distortion measure (5.15) using embedding operations of three different cardinalities.

More formaly, let $\mathbf{x} \in \{0,\dots,255\}^{n_1 \times n_2}$ be an $n_1 \times n_2$ grayscale cover image, $n = n_1 n_2$, represented in the spatial domain. Define the co-occurrence matrix computed from horizontal pixel differences $D^{\rightarrow}_{i,j}(\mathbf{x}) = x_{i,j+1} - x_{i,j}$, $i = 1,\dots,n_1$, $j = 1,\dots,n_2 - 1$:

$$A^{\rightarrow}_{p,q,r}(\mathbf{x}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2-3} \frac{[(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1}, D^{\rightarrow}_{i,j+2})(\mathbf{x}) = (p,q,r)]}{n_1(n_2-3)},$$

where $[(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1}, D^{\rightarrow}_{i,j+2})(\mathbf{x}) = (p,q,r)] = [(D^{\rightarrow}_{i,j}(\mathbf{x}) = p) \& (D^{\rightarrow}_{i,j+1}(\mathbf{x}) = q) \& (D^{\rightarrow}_{i,j+2}(\mathbf{x}) = r)]$. Clearly, $A^{\rightarrow}_{p,q,r}(\mathbf{x}) \in [0,1]$ is the normalized count of neighboring quadruples of pixels $\{x_{i,j}, x_{i,j+1}, x_{i,j+2}, x_{i,j+3}\}$ with differences $x_{i,j+1} - x_{i,j} = p$, $x_{i,j+2} - x_{i,j+1} = q$, and $x_{i,j+3} - x_{i,j+2} = r$ in the entire image. The superscript arrow "$\rightarrow$" denotes the fact that the differences are computed by subtracting the left pixel from the right one. The matrices $A^{\nearrow}_{p,q,r}(\mathbf{x})$, $A^{\uparrow}_{p,q,r}(\mathbf{x})$, and $A^{\nwarrow}_{p,q,r}(\mathbf{x})$ are defined similarly. Let $y_{i,j}\mathbf{x}_{\sim i,j}$ be an image obtained from $\mathbf{x}$ by replacing the (i,j)th pixel with value $y_{i,j}$. Finally, the PI defines the distortion measure $D(\mathbf{y}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \rho_{i,j}(y_{i,j})$ by

$$\rho_{i,j}(y_{i,j}) = \sum_{\substack{p,q,r \in \{-255,\dots,255\} \\ s \in \{\rightarrow,\nearrow,\uparrow,\nwarrow\}}} w_{p,q,r} |A^{s}_{p,q,r}(\mathbf{x}) - A^{s}_{p,q,r}(y_{i,j}\mathbf{x}_{\sim i,j})|, \tag{5.15}$$

where $w_{p,q,r} = 1/(1+\sqrt{p^2+q^2+r^2})$ are heuristically chosen weights.

[Steganalysis setup and experimental results]

All tests were carried out on the BOWS2 database [2] containing approximately 10,800 grayscale images with a fixed size of 512×512 pixels coming from rescaled and cropped natural images of various sizes. Steganalysis was implemented using the second-order SPAM feature set with $T = 3$ [64].

Figure 5.12 contains the comparison of embedding algorithms implementing the PLS and DLS with the costs (5.15). All algorithms are contrasted with LSB matching simulated on the binary and ternary bounds. To compare the effect of practical codes, the PI first simulated the embedding algorithm as if the best codes were available and then compared these results with algorithms implemented using STCs with $h = 10$. Both types of senders are implemented with binary, ternary ($\mathcal{I}_i = \{x_i - 1, \dots, x_i + 1\}$), and pentary ($\mathcal{I}_i = \{x_i - 2, \dots, x_i + 2\}$) embedding operations. Before embedding, the binary embedding operation was initialized to $\mathcal{I}_i = \{x_i, y_i\}$ with y_i randomly chosen from $\{x_i - 1, x_i + 1\}$. The reported payload for the DLS with a fixed D_ϵ was calculated as an average over the whole database after embedding.

The relative horizontal distance between the corresponding dashed and solid lines in Figure 5.12 is bounded by the coding loss. Most of the proposed algorithms are undetectable for relative payloads $\alpha \leq 0.2$ bits per pixel (bpp). For payloads $\alpha \leq 0.5$, the DLS is more secure. For larger payloads, the distortion measure seems to fail to capture the statistical detectability correctly and thus the algorithms are more detectable than when implemented in the payload-limited regime. Finally, the results suggest that larger embedding changes are useful for steganography when placed adaptively.

5.7. **Discussion.** The concept of embedding in steganography that minimizes a distortion function is connected to many basic principles used for constructing embedding schemes for complex cover sources today, including the principle of minimal-embedding-impact [37], approximate model-preservation [65], or the Gibbs construction [23]. The current work describes a complete practical framework for constructing steganographic schemes that embed by minimizing an additive distortion function. Once the steganographer specifies the form of the distortion function, the proposed framework provides all essential tools for constructing practical embedding schemes working close to their theoretical bounds. The methods are not limited to binary embedding operations and allow the embedder to choose the amplitude of embedding changes dynamically based on the cover-image content. The distortion function or the embedding operation do not need to be shared with the recipient. In fact, they can even change from image to image. The framework can be thought of as an off-the-shelf method that allows practitioners to concentrate on the problem of designing the distortion measure instead of the problem of how to construct practical embedding schemes.

The merit of the proposed algorithms is demonstrated experimentally by implementing them for the JPEG and spatial domains and showing an improvement in statistical detectability as measured by state-of-the-art blind steganalyzers. The PI has demonstrated that larger embedding changes provide a significant gain in security when placed adaptively. Finally, the construction is not limited to embedding with larger amplitudes but can be used, e.g., for embedding in color images, where the LSBs of all three colors can be seen as 3-bit symbols on which the cost functions are defined. Applications outside the scope of digital images are possible as long as the costs can be defined.

The embedding using an additive distortion function has been greatly extended to essentially arbitrary distortion functions as described in [23] and in the next section.

6. Gibbs Construction in Steganography

This section describes one of the main achievements of this effort – a general construction for building steganographic schemes using the principle of minimum embedding distortion. The contribution is based on a connection between steganography design by minimizing embedding distortion and statistical physics. The unique aspect of this work and one that distinguishes it from prior art is that the distortion function is allowed to be arbitrary, which permits considering spatially-dependent embedding changes. The PI provides a complete theoretical framework and describes practical tools, such as the thermodynamic integration for computing the rate–distortion bound and the Gibbs sampler, for simulating the impact of optimal embedding schemes and constructing practical algorithms. The proposed framework reduces the design of secure steganography in empirical covers to the problem of finding local potentials for the distortion function that correlate with statistical detectability in practice. By working out the proposed methodology in detail for a specific choice of the distortion function, the PI experimentally validates the approach and discusses various options available to the steganographer in practice.

There exist two general and widely used principles for designing steganographic methods for empirical cover objects, such as digital images. The first one is model-preserving steganography in which the designer adopts a model of the cover source and then designs the embedding to either completely or approximately preserve the model [45, 69, 72, 74, 81]. This way, one can provide mathematical guarantee that the embedding is perfectly secure (or ϵ-secure) within the chosen model. A problem is that empirical cover objects are notoriously difficult to model accurately, and, as history testifies, the model mismatch can be exploited by an attacker to construct a sensitive detection scheme. Even worse, preserving an oversimplified model could introduce a security weakness [8, 56, 91]. An obvious remedy is to use more complicated models that would better approximate the cover source. The major obstacle here is that most current model-preserving steganographic constructions are specific to a certain model and do not adapt easily to more complex models.

The second, quite pragmatic, approach avoids modeling the cover source altogether and, instead, minimizes a heuristically-defined embedding distortion (impact). Matrix embedding [15], wet paper codes [36], and minimal embedding distortion steganography [26, 31, 33, 55, 71] are examples of this philosophy. Despite its heuristic nature, the principle of minimum embedding distortion has produced the most secure steganographic methods for digital media known today, at least in terms of low statistical detectability as measured using blind steganalyzers [37, 55, 58, 71]. Most of these schemes, however, use a distortion function that is additive – the total distortion is a sum of individual pixel distortions *computed from the cover image*. Fundamentally, such a distortion function cannot capture interactions among embedding changes, which leads to suboptimality in practice. This deficiency affects especially adaptive schemes for which the embedding changes have a tendency to form clusters because the pixel distortion is derived from local content or some content-dependent side-information. For example, the embedding changes might follow edges or be concentrated in textured regions.

One discovers a relationship between both embedding principles when the distortion function is defined as a weighted norm of the difference between feature vectors of cover and stego objects in some properly chosen feature space [56, 66], an example of which are spaces utilized by blind steganalyzers. The projection onto the feature space is essentially equivalent to modeling the objects in a lower-dimensional Euclidean space. Consequently, minimizing the distortion between cover and stego objects in the feature space now becomes closely tied to model preservation. Yet again, in this case the distortion cannot be written as a sum of individual pixel distortions also because the features contain higher-order statistics, such as sample transition probability matrices of pixels or DCT coefficients modeled as Markov chains [11, 64, 67, 77].

The importance of modeling interactions among embedding changes in steganography has been indirectly recognized by the designers of MPSteg [10] (Matching Pursuit Steganography) and YASS [73, 80]. In MPSteg, the authors use an overcomplete basis and embed messages by replacing small blocks with other blocks with the hope of preserving dependencies among neighboring pixels. The YASS algorithm taught us that a high embedding distortion may not directly manifest as a high statistical detectability, a curious property that can most likely be attributed to the fact that the embedding modifications are content driven and mutually correlated. Both approaches are heuristic in nature and leave many important issues unanswered, including establishing performance bounds, evaluating the methods' performance w.r.t. to these bounds, and creating a methodology for achieving near-optimal performance.

The above discussion underlines the need for a more systematic approach to steganography that can consider mutual interaction of embedding modifications, which is the topic of this section. The main contribution is a general framework for embedding using arbitrary distortion functions and a complete practical methodology for minimizing embedding distortion in steganography. The approach is flexible as well as modular and allows the steganographer to work with non-additive distortion functions. The PI provides algorithms for computing the proper theoretical bounds expressing the maximal payload embeddable with a bounded distortion, for simulating the impact of a stegosystem operating on the bound, and for designing practical steganographic algorithms that operate near the bound. The algorithms leverage standard tools used in statistical physics, such as Markov chain Monte Carlo samplers or the thermodynamic integration.

In the next section, the PI recalls the basic result that embedding changes made by a steganographic method that minimizes embedding distortion must follow a particular form of Gibbs distribution. The main purpose of this section is to establish terminology and make connections between the concepts used in steganography and those in statistical physics. In Section 6.2, the PI introduces the so-called separation principle, which includes several distinct tasks that must be addressed when developing a practical steganographic method. In particular, to design and evaluate practical schemes one needs to establish the relationship between the maximal payload embeddable using bounded distortion (the rate–distortion bound) and be able to simulate the impact of a scheme operating on the bound. In the special case when the embedding distortion can be expressed as a sum of distortions at individual pixels computed from the cover image (the so-called non-interacting embedding changes), the design of near-optimal embedding algorithms has been successfully resolved in the past. For completeness, and because the proposed framework builds upon these results, the PI briefly summarizes such known achievements in Section 6.3. Continuing with the case of a general distortion function, in Section 6.4 the PI describes two useful tools for steganographers – the Gibbs sampler and the thermodynamic integration. The Gibbs sampler can be used to simulate the impact of optimal embedding and to construct practical steganographic schemes (in Sections 6.5 and 6.6). The thermodynamic integration is a method for estimating the entropy and partition function in statistical physics and it is used for computing the rate–distortion bound in steganography. The design of practical embedding schemes begins in Section 6.5 with the study of distortion functions that can be written as a sum of local potentials defined on cliques. In Section 6.6, the PI first discusses various options the new framework offers to the steganography designer and then makes a connection between local potentials and image models used in blind steganalysis. The proposed framework is experimentally validated in Section 6.7 with a discussion of various implementation issues. Finally, the section is concluded in Section 6.8.

6.1. **Gibbs distribution minimizes embedding distortion.** The PI first recalls a well-known and quite general fact that, for a given expected embedding distortion, the maximal payload is embedded when the embedding changes follow a Gibbs distribution. This establishes a connection between steganography and statistical physics, which, later in this section, will allow computing rate–distortion bounds, simulate the impact of optimal embedding, and construct practical embedding algorithms.

Although the entire framework is certainly applicable to steganography in other objects than digital images, it is described using the terms "image" and "pixel" for concreteness to simplify the language and to allow a smooth transition from theory to experimental validation, which is carried out for digital images. An image $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X} \triangleq \mathcal{I}^n$ is a regular lattice of elements (pixels) $x_i \in \mathcal{I}$, $i \in \mathcal{S}$, $\mathcal{S} = \{1, \ldots, n\}$. The dynamic range, $\mathcal{I}$, depends on the character of the image data. For example, for an 8-bit grayscale image, $\mathcal{I} = \{0, 1, \ldots, 255\}$. In general, x_i can stand not only for light intensity values in a raster image but also for transform coefficients, palette indices, audio samples, etc. The proposed framework remains valid even when x_i is organized into an arbitrary graph structure. For notational simplicity and convenience, additional conventions are adopted. Given $\mathcal{J} \subset \mathcal{S}$, $\mathbf{x}_{\mathcal{J}} \triangleq \{x_i | i \in \mathcal{J}\}$ and $\mathbf{x}_{\sim\mathcal{J}} \triangleq \{x_i | i \in \mathcal{S} - \mathcal{J}\}$. The image $(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)$ will be abbreviated as $y_i\mathbf{x}_{\sim i}$.

Every steganographic embedding scheme considered here will be associated with a mapping that assigns to each cover $\mathbf{x} \in \mathcal{X}$ the pair $\{\mathcal{Y}, \pi\}$. Here, $\mathcal{Y} \subset \mathcal{X}$ is the set of all stego images $\mathbf{y}$ into which $\mathbf{x}$ is allowed to be modified by embedding and π is a probability mass function on $\mathcal{Y}$ that characterizes the actions of the sender. The embedding algorithm is such that, for a given cover $\mathbf{x}$, the stego image $\mathbf{y} \in \mathcal{Y}$ is sent with probability $\pi(\mathbf{y})$. The stego image is thus a random variable $\mathbf{Y}$ over $\mathcal{Y}$ with the distribution $P(\mathbf{Y} = \mathbf{y}) = \pi(\mathbf{y})$. Technically, the set $\mathcal{Y}$ and all concepts derived from it in this report depend on $\mathbf{x}$. However, because $\mathbf{x}$ is simply a parameter that one *fixes in the very beginning*, the notation is simplified by not making the dependence on $\mathbf{x}$ explicit. Finally, note that the maximal expected payload that the sender can communicate to the receiver in this manner is the entropy

$$H(\pi) \triangleq H(\mathbf{Y}) = -\sum_{\mathbf{y} \in \mathcal{Y}} \pi(\mathbf{y}) \log \pi(\mathbf{y}). \tag{6.1}$$

To put it another way, the PI defines a steganographic method from the point of view of how it modifies the cover and only then deals with the issues of how to use it for communication and how to optimize its performance. The optimization will involve finding the distribution π for given $\mathbf{x}$, $\mathcal{Y}$, and payload (distortion).

The following special form of the set $\mathcal{Y}$: $\mathcal{Y} = \mathcal{I}_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n$ will be considered, where $\mathcal{I}_i \subset \mathcal{I}$. For example, in Least Significant Bit (LSB) embedding, $\mathcal{I}_i = \{x_i, \overline{x}_i\}$, where the bar denotes the operation of flipping the LSB. In LSB matching [49] (also called ± 1 embedding) in an 8-bit grayscale image $\mathbf{x}$, $\mathcal{I}_i = \{x_i - 1, x_i, x_i + 1\}$ whenever $x_i \notin \{0, 255\}$ and $\mathcal{I}_i$ is appropriately modified for the boundary cases. When $|\mathcal{I}_i| = 2$ or 3 for all i, one speaks of binary and ternary embedding, respectively. In general, however, the size of every set $\mathcal{I}_i$ can be different. For example, pixels not allowed to be modified during embedding (the so-called wet pixels [36]) have $\mathcal{I}_i = \{x_i\}$.

By sending a slightly modified version $\mathbf{y}$ of the cover $\mathbf{x}$, the sender introduces a distortion, which will be measured using a distortion function

$$D : \mathcal{Y} \to \mathbb{R}, \tag{6.2}$$

that is bounded, i.e., $|D(\mathbf{y})| < K$, for all $\mathbf{y} \in \mathcal{Y}$ for some sufficiently large K. Note that D also depends on $\mathbf{x}$. Allowing the distortion to be negative does not cause any problems because an embedding algorithm minimizes D if and only if it minimizes the non-negative distortion $D+K$. The need for negative distortion will become apparent later in Section 6.5.1.

The expected embedding distortion introduced by the sender is

$$E_\pi[D] = \sum_{\mathbf{y}\in\mathcal{Y}} \pi(\mathbf{y})D(\mathbf{y}). \tag{6.3}$$

An important premise made now is that the sender is able to define the distortion function so that it is related to statistical detectability.[18] This assumption is motivated by a rather large body of experimental evidence, such as [37, 58], that indicates that even simple distortion measures that merely count the number of embedding changes correlate well with statistical detectability in the form of decision error of steganalyzers trained on cover and stego images. In general, steganographic methods that introduce smaller distortion disturb the cover source less than methods that embed with larger distortion.

Distortion-limited sender. To maximize the security, the so-called distortion-limited sender attempts to find a distribution π on $\mathcal{Y}$ that has the highest entropy and whose expected embedding distortion does not exceed a given D_ϵ:

$$\underset{\pi}{\text{maximize}} \quad H(\pi) = -\sum_{\mathbf{y}\in\mathcal{Y}} \pi(\mathbf{y}) \log \pi(\mathbf{y}) \tag{6.4}$$

$$\text{subject to} \quad E_\pi[D] = \sum_{\mathbf{y}\in\mathcal{Y}} \pi(\mathbf{y})D(\mathbf{y}) = D_\epsilon. \tag{6.5}$$

By fixing the distortion, the sender fixes the security and aims to communicate as large payload as possible at this level of security. The maximization in (6.4) is carried over all distributions π on $\mathcal{Y}$. The PI comments on whether the distortion constraint should be in the form of equality or inequality shortly.

Payload-limited sender. Alternatively, in practice it may be more meaningful to consider the payload-limited sender who faces a complementary task of embedding a *given* payload of m bits with minimal possible distortion. The optimization problem is to determine a distribution π that communicates a required payload while minimizing the distortion:

$$\underset{\pi}{\text{minimize}} \quad E_\pi[D] = \sum_{\mathbf{y}\in\mathcal{Y}} \pi(\mathbf{y})D(\mathbf{y}) \tag{6.6}$$

$$\text{subject to} \quad H(\pi) = m. \tag{6.7}$$

The optimal distribution π for both problems has the Gibbs form

$$\pi_\lambda(\mathbf{y}) = \frac{1}{Z(\lambda)} \exp(-\lambda D(\mathbf{y})), \tag{6.8}$$

where $Z(\lambda)$ is the normalizing factor

$$Z(\lambda) = \sum_{\mathbf{y}\in\mathcal{Y}} \exp(-\lambda D(\mathbf{y})). \tag{6.9}$$

The optimality of π_λ follows immediately from the fact that for any distribution μ with $E_\mu[D] = \sum_{\mathbf{y}\in\mathcal{Y}} \mu(\mathbf{y})D(\mathbf{y}) = D_\epsilon$, the difference between their entropies, $H(\pi_\lambda) - H(\mu) = D_{\mathrm{KL}}(\mu||\pi_\lambda) \geq 0$ [92]. The scalar parameter $\lambda > 0$ needs to be determined from the distortion constraint (6.5) or from the payload constraint (6.7), depending on the type of the sender. Provided m or D_ϵ are in the feasibility region of their corresponding constraints, the value of λ is unique. This follows from the fact that both the expected distortion and the entropy are monotone decreasing in λ. To see this, realize that by direct evaluation

$$\frac{\partial}{\partial\lambda} E_{\pi_\lambda}[D] = -Var_{\pi_\lambda}[D] \leq 0, \tag{6.10}$$

where $Var_{\pi_\lambda}[D] = E_{\pi_\lambda}[D^2] - (E_{\pi_\lambda}[D])^2$. Substituting (6.8) into (6.1), the entropy of the Gibbs distribution can be written as

[18] The ability of a warden to distinguish between cover and stego images using statistical hypothesis testing.

$$H(\pi_\lambda) = \log Z(\lambda) + \frac{1}{\ln 2}\lambda E_{\pi_\lambda}[D]. \tag{6.11}$$

Upon differentiating and using (6.10), one obtains

$$\frac{\partial}{\partial\lambda}H(\pi_\lambda) = \frac{1}{\ln 2}\left(\frac{Z'(\lambda)}{Z(\lambda)} + E_{\pi_\lambda}[D] - \lambda Var_{\pi_\lambda}[D]\right) \tag{6.12}$$

$$= -\frac{\lambda}{\ln 2}Var_{\pi_\lambda}[D] \le 0. \tag{6.13}$$

The monotonicity also means that the equality distortion constraint in the optimization problem (6.5) can be replaced with inequality, which is perhaps more appropriate given the motivating discussion above.

By varying $\lambda \in [0,\infty)$, one obtains a relationship between the maximal expected payload (6.1) and the expected embedding distortion (6.3). For brevity, this relationship will be called the rate–distortion bound. What distinguishes this concept from a similar notion defined in information theory is that the bound is considered for a *given* cover $\mathbf{x}$ rather than for $\mathbf{X}$, which is a random variable. At this point, it is appropriate to note that while it is certainly possible to consider $\mathbf{x}$ to be generated by a cover source with a known distribution and approach the design of steganography from a different point of view, namely one in which π_λ is determined by minimizing the KL divergence between the distributions of cover and stego images while satisfying a payload constraint, it is not done here.

Finally, note that the assumption $|D(\mathbf{y})| < K$ implies that all stego objects appear with nonzero probability, $\pi_\lambda(\mathbf{y}) \ge \frac{1}{Z(\lambda)}\exp(-\lambda K)$, a fact that is crucial for the theory developed here.

Remark 12. In statistical physics, the term distortion is known as energy. The optimality of Gibbs distribution is formulated as the Gibbs variational principle: "Among all distributions with a given energy, the Gibbs distribution (6.8) has the highest entropy." The parameter λ is called the inverse temperature, $\lambda = 1/kT$, where T is the temperature and k the Boltzmann constant. The normalizing factor $Z(\lambda)$ is called the partition function.

6.2. **The separation principle.** The design of steganographic methods that attempt to minimize embedding distortion should be driven by their performance. The obvious choice here is to contrast the performance with the rate–distortion bound. This is a meaningful comparison for the distortion-limited sender who can assess the performance of a practical embedding scheme by its loss of payload w.r.t. the maximum payload embeddable using a fixed distortion. This so-called "coding loss" informs the sender of how much payload is lost for a fixed statistical detectability. On the other hand, it is much harder for the payload-limited sender to assess how the increased distortion of a suboptimal practical scheme impacts statistical detectability in practice. One could resolve this rather important practical issue if one was able to simulate the impact of a scheme that operates *on the bound*.[19] Because the problems of establishing the bounds, simulating optimal embedding, and creating a practical embedding algorithm are really three separate problems, this reasoning will be called the *separation principle*. It involves addressing the following three tasks:

(1) **Establishing the rate–distortion bounds.** This means solving the optimization problems (6.4) or (6.6) and expressing the largest payload embeddable using a bounded distortion (or minimal distortion needed to embed a given payload). These bounds inform the steganographer about the best performance that can be theoretically achieved. Depending on the form of the distortion function D, establishing the bounds is usually rather challenging and one may have to resort to numerical methods (Section 6.4.2). For an additive distortion (to be precisely defined shortly), an analytic form of the bounds may be obtained (Section 6.3).

(2) **Simulating an optimal embedding method.** Often, it is very hard to construct a practical embedding method that performs close to the bound. However, one may be able to simulate the impact of such an optimal method and thus subject it to tests using steganalyzers even when it is not known how to construct a practical embedding algorithm or even compute the bound (see Section 6.4). This is important for developers as one can effectively "prune" the design process and focus on implementing the most promising candidates. The simulator will also inform the payload-limited sender about the potential improvement in statistical undetectability should the theoretical performance gap be closed. A simple example is provided by the case of the Hamming distortion function $D(\mathbf{y}) = \sum_i [y_i \ne x_i]$. Here, the maximal relative payload $\alpha = m/n$ (in bits per pixel or bpp) is bounded by $\alpha \le h(\beta)$, where $\beta = \frac{1}{n}D_\epsilon$ is the relative embedding distortion known as the change rate. In this case, one can simulate the embedding impact of the optimal scheme by independently changing each pixel with probability $h^{-1}(\alpha)$.

[19] A scheme whose embedding distortion and payload lay on the rate–distortion bound derived for a given cover.

(3) **Constructing a practical near-optimal embedding method.** This point is of most interest to practitioners. The bounds and the simulator are necessary to evaluate the performance of any practical scheme. The designer tries to maximize the embedding throughput (the number of bits embedded per unit time) while embedding as close to the distortion bound as possible.

It should be stressed at this point that even though the optimal distribution of embedding modifications has a known analytic expression (6.8), it may be infeasible to compute the individual probabilities $\pi_\lambda(\mathbf{y})$ due to the complexity of evaluating the partition function $Z(\lambda)$, which is a sum over all $\mathbf{y}$, whose count can be a very large number even for small images. (For example, there are 2^n binary flipping patterns in LSB embedding.) This also implies that at present it is not clear how to compute the expected distortion (6.3) or the entropy (6.1) (these tasks are postponed to Section 6.4). Fortunately, in many cases of practical interest it is not necessary to evaluate $\pi_\lambda(\mathbf{y})$ and it will be enough to merely *sample from* π_λ. The ability to sample from π_λ is sufficient to simulate optimal embedding and realize practical embedding algorithms, and, in this case, even compute the rate–distortion bound.

In some special cases, however, such as when the embedding changes do not interact, the distortion D is additive and one can easily compute λ and the probabilities, evaluate the expected distortion and payload, and even construct near-optimal embedding schemes. As this special case will be used later in Section 6.6 to design steganography with more general distortion functions D, it is briefly reviewed in the next section.

6.3. **Non-interacting embedding changes.** When the distortion function D is additive over the pixels,

$$D(\mathbf{y}) = \sum_{i=1}^{n} \rho_i(y_i), \tag{6.14}$$

with bounded $\rho_i : \mathcal{I}_i \to \mathbb{R}$, the embedding changes are said to not interact. In this case, the probability $\pi_\lambda(\mathbf{y})$ can be factorized into a product of marginal probabilities of changing the individual pixels (this follows directly from (6.8)):

$$\pi_\lambda(\mathbf{y}) = \prod_{i=1}^{n} \pi_\lambda(y_i) = \prod_{i=1}^{n} \frac{\exp(-\lambda\rho_i(y_i))}{\sum_{t_i\in\mathcal{I}_i}\exp(-\lambda\rho_i(t_i))}. \tag{6.15}$$

The expected distortion and the maximal payload are:

$$E_{\pi_\lambda}[D] = \sum_{i=1}^{n}\sum_{t_i\in\mathcal{I}_i} \pi_\lambda(t_i)\rho_i(t_i), \tag{6.16}$$

$$H(\pi_\lambda) = -\sum_{i=1}^{n}\sum_{t_i\in\mathcal{I}_i} \pi_\lambda(t_i)\log\pi_\lambda(t_i). \tag{6.17}$$

The impact of optimal embedding can be simulated by changing x_i to y_i with probabilities $\pi_\lambda(y_i)$ independently of the changes at other pixels. Since these probabilities can now be easily evaluated for a fixed λ, finding λ that satisfies the distortion ($E_{\pi_\lambda}[D] = D_\epsilon$) or the payload ($H(\pi_\lambda) = m$) constraint amounts to solving an algebraic equation for λ (see [31] or [29]). Because both the expected distortion and the entropy are monotone w.r.t. λ, the solution is unique. The only practical near-optimal embedding algorithm for this case known to the PI is based on syndrome-trellis codes [25].

It will be instructional to work out as an example the details of the special case of binary embedding for which $\mathcal{I}_i = \{x_i^{(0)}, x_i^{(1)}\}$ with $x_i^{(0)} = x_i$. Thus, ρ_i attains only two values, $\rho_i^{(t)} = \rho_i(x_i^{(t)})$, $t = 0, 1$. The PI stresses at this point that it is *not* assumed that $\rho_i^{(0)} = 0$ or even that $\rho_i^{(1)} \geq \rho_i^{(0)}$. This fact will be important when implementing practical embedding schemes in Section 6.5.1. The above expressions simplify to

$$\pi_\lambda\big(x_i^{(1)}\big) = \frac{\exp\big(-\lambda\rho_i^{(1)}\big)}{\exp\big(-\lambda\rho_i^{(1)}\big) + \exp\big(-\lambda\rho_i^{(0)}\big)} \tag{6.18}$$

$$= \frac{1}{1+\exp\big(-\lambda\big(\rho_i^{(0)} - \rho_i^{(1)}\big)\big)} \triangleq p_i(\lambda), \tag{6.19}$$

$$E_{\pi_\lambda}[D] = \sum_{i=1}^{n} \rho_i^{(0)}(1-p_i(\lambda)) + \rho_i^{(1)}p_i(\lambda), \tag{6.20}$$

$$H(\pi_\lambda) = \sum_{i=1}^{n} h(p_i(\lambda)). \tag{6.21}$$

The smallest distortion any binary embedding algorithm can impose is $D_{\min} = \sum_{i=1}^{n} \min\{\rho_i^{(0)}, \rho_i^{(1)}\}$, which would be incurred when selecting $y_i = x_i^{(t_i)}$, where $t_i = \arg\min_t \{\rho_i^{(t)}\}$. Thus,

$$D(\mathbf{y}) = \sum_{i=1}^{n} \rho_i^{(0)}\left[y_i = x_i^{(0)}\right] + \rho_i^{(1)}\left[y_i = x_i^{(1)}\right] \tag{6.22}$$

$$= D_{\min} + \sum_{i=1}^{n} \varrho_i\left[y_i \neq x_i^{(t_i)}\right], \tag{6.23}$$

where $\varrho_i = |\rho_i^{(1)} - \rho_i^{(0)}|$ is now a vector of non-negative distortions, which allows applying the practical embedding algorithm described in [26]. It accepts on its input a bit stream $\mathbf{c} = (c_1(\mathbf{x}), \ldots, c_n(\mathbf{x}))$ (representing the cover $\mathbf{x}$), the vector of non-negative distortions $(\varrho_1, \ldots, \varrho_n)$, and a binary message. It outputs a modified (stego) bit stream $\mathbf{y} \in \{0,1\}^n$ that conveys the message as a syndrome of a suitably chosen syndrome-trellis code so that the total embedding distortion $\sum_{i=1}^{n} \varrho_i[y_i \neq c_i]$ is near minimal. It follows from (6.23) that binary embedding as defined in this section can be implemented in practice by applying this algorithm to the bit stream $c_i(\tilde{\mathbf{x}})$, $\tilde{\mathbf{x}} = (x_1^{(t_1)}, \ldots, x_n^{(t_n)})$.

The complete derivation of the rate-distortion bound for binary embedding appears, e.g., in Chapter 7 of [29].

6.4. **Simulated embedding and rate–distortion bound.** In Section 6.1, it has been showed that minimal-embedding-distortion steganography should select the stego image $\mathbf{y}$ with probability $\pi_\lambda(\mathbf{y}) \propto \exp(-\lambda D(\mathbf{y}))$ expressed in the form of a Gibbs distribution. The PI now explains a general iterative procedure using which one can sample from any Gibbs distribution and thus simulate optimal embedding. The method is recognized as one of the Markov Chain Monte Carlo (MCMC) algorithms known as the Gibbs sampler.[20] This sampling algorithm will allow constructing practical embedding schemes in Sections 6.5 and 6.6. It will also be explained how to compute the rate–distortion bound for a fixed image using the thermodynamic integration. The Gibbs sampler and the thermodynamic integration appear, for example, in [92] and [62], respectively.

6.4.1. *The Gibbs sampler.* The PI starts with defining the local characteristics of a Gibbs field as the conditional probabilities of the ith pixel attaining the value y_i' conditioned on the rest of the image:

$$\pi_\lambda(Y_i = y_i' | \mathbf{Y}_{\sim i} = \mathbf{y}_{\sim i}) = \frac{\pi_\lambda(y_i' \mathbf{y}_{\sim i})}{\sum_{t_i \in \mathcal{I}_i} \pi_\lambda(t_i \mathbf{y}_{\sim i})}. \tag{6.24}$$

For all possible stego images $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$, the local characteristics (6.24) define the following matrices $\mathbb{P}(i)$, for each pixel $i \in \{1, \ldots, n\}$:

$$P_{\mathbf{y},\mathbf{y}'}(i) = \begin{cases} \pi_\lambda(Y_i = y_i' | \mathbf{Y}_{\sim i} = \mathbf{y}_{\sim i}) & \text{when } \mathbf{y}'_{\sim i} = \mathbf{y}_{\sim i} \\ 0 & \text{otherwise.} \end{cases} \tag{6.25}$$

Every matrix $\mathbb{P}(i)$ has $|\mathcal{Y}|$ rows and the same number of columns (which means it is very large) and its elements are mostly zero except when $\mathbf{y}'$ was obtained from $\mathbf{y}$ by modifying y_i to y_i' and all other pixels stayed the same. Because $\mathbb{P}(i)$ is stochastic (the sum of its rows is one),

$$\sum_{\mathbf{y}' \in \mathcal{Y}} P_{\mathbf{y},\mathbf{y}'}(i) = 1, \text{ for all rows } \mathbf{y}, \tag{6.26}$$

$\mathbb{P}(i)$ is a transition probability matrix of some Markov chain on $\mathcal{Y}$. All such matrices satisfy the so-called detailed balance equation

$$\pi_\lambda(\mathbf{y}) P_{\mathbf{y},\mathbf{y}'}(i) = \pi_\lambda(\mathbf{y}') P_{\mathbf{y}',\mathbf{y}}(i), \quad \text{for all } \mathbf{y}, \mathbf{y}' \in \mathcal{Y},\ i. \tag{6.27}$$

[20] More detailed discussion regarding our choice of the MCMC sampler appear later in this section.

Algorithm 1 One sweep of a Gibbs sampler.

1: Set pixel counter $i = 1$
2: **while** $i \leq n$ **do**
3: Compute the local characteristics:

$$P_{y'_{\sigma(i)}\mathbf{y}_{\sim\sigma(i)},\mathbf{y}}(\sigma(i)),\ y'_{\sigma(i)} \in \mathcal{I}_{\sigma(i)} \tag{6.34}$$

4: Select one $y'_{\sigma(i)} \in \mathcal{I}_{\sigma(i)}$ pseudorandomly according to the probabilities (6.34) and change $y_{\sigma(i)} \leftarrow y'_{\sigma(i)}$
5: $i \leftarrow i + 1$
6: **end while**
7: **return** $\mathbf{y}$

To see this, realize that unless $\mathbf{y}_{\sim i} = \mathbf{y}'_{\sim i}$, one is looking at the trivial equality $0 = 0$. For $\mathbf{y}_{\sim i} = \mathbf{y}'_{\sim i}$, one obtains the following chain of equalities:

$$\pi_\lambda(\mathbf{y})P_{\mathbf{y},\mathbf{y}'}(i) \overset{(a)}{=} \pi_\lambda(\mathbf{y})\frac{\pi_\lambda(y'_i\mathbf{y}_{\sim i})}{\sum_{t_i\in\mathcal{I}_i}\pi_\lambda(t_i\mathbf{y}_{\sim i})} \tag{6.28}$$

$$\overset{(b)}{=} \frac{\pi_\lambda(\mathbf{y})\pi_\lambda(\mathbf{y}')}{\sum_{t_i\in\mathcal{I}_i}\pi_\lambda(t_i\mathbf{y}_{\sim i})} \tag{6.29}$$

$$= \pi_\lambda(\mathbf{y}')\frac{\pi_\lambda(\mathbf{y})}{\sum_{t_i\in\mathcal{I}_i}\pi_\lambda(t_i\mathbf{y}'_{\sim i})} \tag{6.30}$$

$$\overset{(c)}{=} \pi_\lambda(\mathbf{y}')P_{\mathbf{y}',\mathbf{y}}(i). \tag{6.31}$$

Equality (a) follows from the definition of $\mathbb{P}(i)$ (6.25), (b) from the fact that $\mathbf{y}_{\sim i} = \mathbf{y}'_{\sim i}$, and (c) from $\pi_\lambda(\mathbf{y}) = \pi_\lambda(y_i\mathbf{y}'_{\sim i})$ and again (6.25).

Next, the PI defines the boldface symbol $\boldsymbol{\pi}_\lambda \in [0,\infty)^{|\mathcal{Y}|}$ as the vector of $|\mathcal{Y}|$ non-negative elements $\pi_\lambda = \pi_\lambda(\mathbf{y})$, $\mathbf{y} \in \mathcal{Y}$. Using (6.27) and then (6.26), one can now easily show that the vector $\boldsymbol{\pi}_\lambda$ is the left eigenvector of $\mathbb{P}(i)$ corresponding to the unit eigenvalue:

$$(\boldsymbol{\pi}_\lambda\mathbb{P}(i))_{\mathbf{y}'} = \sum_{\mathbf{y}\in\mathcal{Y}}\pi_\lambda(\mathbf{y})P_{\mathbf{y},\mathbf{y}'}(i) \tag{6.32}$$

$$= \sum_{\mathbf{y}\in\mathcal{Y}}\pi_\lambda(\mathbf{y}')P_{\mathbf{y}',\mathbf{y}}(i) = \pi_\lambda(\mathbf{y}'). \tag{6.33}$$

In (6.32), $(\boldsymbol{\pi}_\lambda\mathbb{P}(i))_{\mathbf{y}'}$ is the $\mathbf{y}'$th element of the product of the vector $\boldsymbol{\pi}_\lambda$ and the matrix $\mathbb{P}(i)$.

It is now time to describe the Gibbs sampler [40], which is a key element in the framework. Let σ be a permutation of the index set $\mathcal{S}$ called the visiting schedule ($\sigma(i)$, $i = 1,\ldots,n$ is the ith element of the permutation σ). One sample from $\boldsymbol{\pi}_\lambda$ is then obtained by repeating a series of "sweeps" defined below. As the sweeps and the Gibbs sampler are explained, the reader is advised to inspect Algorithm 1 to better understand the process.

The sampler is initialized by setting $\mathbf{y}$ to some initial value. For faster convergence, a good choice is to select y_i from $\mathcal{I}_i$ according to the local characteristics $\pi_\lambda(y_i\mathbf{x}_{\sim i})$. A sweep is a procedure applied to an image during which all pixels are updated sequentially in the order defined by the visiting schedule σ. The pixels are updated based on their local characteristics (6.24) computed from the current values of the stego image $\mathbf{y}$. The entire sweep can be described by a transition probability matrix $\mathbb{P}(\sigma)$ obtained by matrix-multiplications of the individual transition probability matrices $\mathbb{P}(\sigma(i))$:

$$P_{\mathbf{y},\mathbf{y}'}(\sigma) \triangleq (\mathbb{P}(\sigma(1))\cdot\mathbb{P}(\sigma(2))\cdots\mathbb{P}(\sigma(n)))_{\mathbf{y},\mathbf{y}'}. \tag{6.35}$$

After each sweep, the next sweep continues with the current image $\mathbf{y}$ as its starting position. It should be clear from the algorithm that at the end of each sweep each pixel i has a non-zero probability to get into any of its states from $\mathcal{I}_i$ defined by the embedding operation (because D is bounded). This means that all elements of $\mathcal{Y}$ will be visited with positive probability and thus the transition probability matrix $\mathbb{P}(\sigma)$ corresponds to a homogeneous irreducible Markov process with a *unique* left eigenvector corresponding to a unit eigenvalue (unique stationary distribution). Because $\boldsymbol{\pi}_\lambda$ is a left eigenvector corresponding to a unit eigenvalue for each matrix $\mathbb{P}(i)$, it is also a left eigenvector for $\mathbb{P}(\sigma)$ and thus its stationary distribution due to its uniqueness. A standard result from the theory of Markov chains (see, e.g. Chapter 4 in [92]) states that, for an irreducible Markov chain, no matter what distribution of embedding changes $\boldsymbol{\nu} \in [0,\infty)^{|\mathcal{Y}|}$ one

starts with, and independently of the visiting schedule σ, with increased number of sweeps, k, the distribution of Gibbs samples converges in norm to the stationary distribution π_λ:

$$||\nu\,(\mathbb{P}(\sigma))^k - \pi_\lambda|| \to 0 \text{ with } k \to \infty \tag{6.36}$$

exponentially fast. This means that in practice one can obtain a sample from π_λ after running the Gibbs sampler for a sufficiently long time.[21] The visiting schedule can be randomized in each sweep as long as each pixel has a non-zero probability of being visited, which is a necessary condition for convergence.

6.4.2. *Simulating optimal embedding.* When applied to steganography, the Gibbs sampler allows the sender to simulate the effect of embedding using a scheme that operates on the bound. It is interesting that this can be done for any distortion function D and without knowing the rate–distortion bound. This is because the local characteristics (6.24)

$$\pi_\lambda(Y_i = y_i' | \mathbf{Y}_{\sim i} = \mathbf{y}_{\sim i}) = \frac{\exp(-\lambda D(y_i' \mathbf{y}_{\sim i}))}{\sum_{t_i \in \mathcal{I}_i} \exp(-\lambda D(t_i \mathbf{y}_{\sim i}))}, \tag{6.37}$$

do not require computing the partition function $Z(\lambda)$. One does need to know the parameter λ, though.

For the distortion-limited sender (6.5), the Gibbs sampler could be used directly to determine the proper value of λ in the following manner. For a given λ, it is known (Theorem 5.1.4 in [92]) that

$$\frac{1}{k}\sum_{j=1}^{k} D(\mathbf{y}^{(j)}) \to E_{\pi_\lambda}[D] \text{ as } k \to \infty \tag{6.38}$$

in L_2 and in probability, where $\mathbf{y}^{(j)}$ is the image obtained after the jth sweep of the Gibbs sampler. This requires running the Gibbs sampler and averaging the individual distortions for a sufficiently long time. When only a finite number of sweeps is allowed, the first few images $\mathbf{y}$ should be discarded to allow the Gibbs sampler to converge close enough to π_λ. The value of λ that satisfies $E_{\pi_\lambda}[D] = D_\epsilon$ can be determined, for example, using a binary search over λ.

To find λ for the payload-limited sender (6.4), one needs to evaluate the entropy $H(\pi_\lambda)$, which can be obtained from $E_{\pi_\lambda}[D]$ using the method of thermodynamic integration [62]. From (6.10) and (6.13), one obtains

$$\frac{\partial}{\partial\lambda} H(\pi_\lambda) = \frac{\lambda}{\ln 2}\frac{\partial}{\partial\lambda} E_{\pi_\lambda}[D]. \tag{6.39}$$

Therefore, the entropy can be estimated from $E_{\pi_\lambda}[D]$ by integrating by parts:

$$H(\pi_\lambda) = H(\pi_{\lambda_0}) + \left[\frac{\lambda'}{\ln 2} E_{\pi_{\lambda'}}[D]\right]_{\lambda_0}^{\lambda} - \frac{1}{\ln 2}\int_{\lambda_0}^{\lambda} E_{\pi_{\lambda'}}[D]\mathrm{d}\lambda'. \tag{6.40}$$

The value of λ that satisfies the entropy (payload) constraint can be again obtained using a binary search. Having obtained the expected distortion and the entropy using the Gibbs sampler and the thermodynamic integration, the rate–distortion bound $[H(\pi_\lambda), E_{\pi_\lambda}[D]]$ can be plotted as a curve parametrized by λ.

In practice, one has to be careful when using (6.38), since no practical guidelines exist for determining a sufficient number of sweeps and heuristic criteria are often used [14, 92]. Although the convergence to π_λ is exponential in the number of sweeps, in general a large number of sweeps may be needed to converge close enough. Generally speaking, the stronger the dependencies between embedding changes the more sweeps are needed by the Gibbs sampler. In theory, the convergence of MCMC methods, such as the Gibbs sampler, may also slow down in the vicinity of "phase transitions," which is loosely defined here as sudden changes in the spatial distribution of embedding changes when only slightly changing the payload (or distortion bound).

In experiments reported later in this section, the Gibbs sampler always behaved well and converged fast. This is attributed to the fact that the dependencies among embedding modifications as measured using our distortion functions are rather weak and limited to short distances. The convergence, however, could become an issue for other types of cover sources with different distortion functions. While it is possible to compute the rate–distortion bounds and simulate optimal embedding using other MCMC algorithms, such as the Metropolis-Hastings sampler [92], that may converge faster than the Gibbs sampler and can exhibit a more robust behavior in practice, it is not clear how to adopt these algorithms for practical embedding. This is because all known coding methods in steganography essentially sample from a distribution of independent symbols. Thus, the Gibbs sampler comes out as a natural choice (Section 6.5) because it works by updating individual pixels, which is exactly the effect of embedding using syndrome-trellis codes [25, 26].

[21] The convergence time may vary significantly depending on the Gibbs field at hand.

FIGURE 6.1. The four-element cross-neighborhood and the tessellation of the index set $\mathcal{S}$ into two disjoint sublattices $\mathcal{S}_e$ and $\mathcal{S}_o$.

FIGURE 6.2. All three possible cliques for the cross-neighborhood.

A notable alternative to the Gibbs sampler and the thermodynamic integration for computing the rate–distortion bound is the Wang–Landau algorithm [87] that estimates the so-called density of stego images (density of states in statistical physics), $g(D)$, defined as the number of stego images $\mathbf{y}$ with distortion (energy) D. The partition function (and thus, via (6.11), the entropy) and the expected distortion can be computed from $g(D)$ by numerical integration:

$$Z(\lambda) \doteq \sum_{D\in\mathcal{D}} g(D)\exp(-\lambda D)\Delta, \tag{6.41}$$

$$E_{\pi_\lambda}[D] \doteq \frac{1}{Z(\lambda)}\sum_{D\in\mathcal{D}} Dg(D)\exp(-\lambda D)\Delta, \tag{6.42}$$

where $\mathcal{D} = \{d_1,\dots,d_{n_D}\}$, $d_1 = -K$, $d_{n_D} = K$, $d_i - d_{i-1} = \Delta$ is a set of discrete values into which the dynamic range of D, $[-K,K]$ is quantized.

The PI notes that in general it is not possible to determine ahead of time which method will provide satisfactory performance. In experiments described in Section 6.7, the thermodynamic integration worked very well and provided results identical to the much more complex Wang–Landau algorithm.

Note that computing the rate–distortion bound is not necessary for practical embedding. In Section 6.5, one introduces a special form of the distortion in terms of a sum over local potentials. In this case, both types of optimal senders can be simulated using algorithms that do not need to compute λ in the fashion described above. This is explained in Sections 6.5.1 and 6.5.2.

6.5. **Local distortion function.** Thanks to the Gibbs sampler, one can simulate the impact of embedding that is optimal in the sense of (6.4) and (6.6) without having to construct a specific steganographic scheme. This is important for steganography design as one can test the effect of various design choices and parameters and then implement only the most promising constructs. However, it is rather difficult to design near-optimal schemes for a general $D(\mathbf{y})$. Fortunately, it is possible to give the distortion function a specific form that will allow constructing practical embedding algorithms. It will be assumed that D is a sum of local potentials defined on small groups of pixels called cliques. This local form of the distortion will be still quite general to capture dependencies among embedding changes and it allows construction of a large spectrum of diverse embedding schemes – a topic left for Section 6.6.

First, the PI defines a neighborhood system as a collection of subsets of the index set $\{\eta(i) \subset \mathcal{S} | i = 1,\dots,n\}$ satisfying $i \notin \eta(i), \forall i$ and $i \in \eta(j)$ if and only if $j \in \eta(i)$. The elements of $\eta(i)$ are called neighbors of pixel i. A subset $c \subset \mathcal{S}$ is a clique if each pair of different elements from c are neighbors. The set of all cliques will be denoted $\mathcal{C}$. The PI does not use the calligraphic font for a clique even though it is a set (and thus deviate here from the established convention) to comply with a well established notation used in previous art.

In this section and in Section 6.6, it will be necessary to address pixels by their two-dimensional coordinates. The PI will hus be switching between using the index set $\mathcal{S} = \{1,\dots,n\}$ and its two-dimensional equivalent $\mathcal{S} = \{(i,j) | 1 \le i \le n_1, 1 \le j \le n_2\}$ hoping that it will cause no confusion for the reader.

Example 13. The four-element cross neighborhood of pixel $x_{i,j}$ consisting of $\{x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}\}$ with a proper treatment at the boundary forms a neighborhood system (see Figure 6.1). The cliques contain either a single pixel (one-element) cliques $\{x_{i,j}\}$ or two horizontally or vertically neighboring pixels, $\{x_{i,j}, x_{i,j+1}\}$, $\{x_{i,j}, x_{i+1,j}\}$ (Figure 6.2). No other cliques exist.

FIGURE 6.3. The eight-element neighborhood and the tessellation of the index set $\mathcal{S}$ into four disjoint sublattices marked with four different symbols.

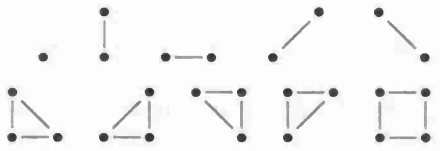

FIGURE 6.4. All possible cliques for the eight-element neighborhood.

Example 14. The eight-element 3×3 neighborhood also forms a neighborhood system (Figure 6.3). The cliques are as in Example 13 as well as all cliques containing pairs of diagonally neighboring pixels, $\{x_{i,j}, x_{i+1,j+1}\}$, $\{x_{i,j}, x_{i-1,j+1}\}$, three-pixel cliques forming a right-angle triangle (e.g., $\{x_{i,j}, x_{i,j+1}, x_{i+1,j}\}$), and four-pixel cliques forming a 2×2 square ($\{x_{i,j}, x_{i,j+1}, x_{i+1,j}, x_{i+1,j+1}\}$) (follow Figure 6.4). No other cliques exist for this neighborhood system.

Each neighborhood system allows tessellation of the index set $\mathcal{S}$ into disjoint subsets (sublattices) whose union is the entire set $\mathcal{S}$, so that any two pixels in each lattice are not neighbors. For example, for the cross-neighborhood $\mathcal{S} = \mathcal{S}_e \cup \mathcal{S}_o$, where

$$\mathcal{S}_e = \{(i,j) | i+j \text{ is even}\}, \qquad \mathcal{S}_o = \{(i,j) | i+j \text{ is odd}\}. \tag{6.43}$$

For the eight-element 3×3 neighborhood, there are four sublattices, $\mathcal{S} = \bigcup_{ab} \mathcal{S}_{ab}$, $1 \le a, b \le 2$, whose structure resembles the Bayer color filter array commonly used in digital cameras [29],

$$\mathcal{S}_{ab} = \{(a+2k, b+2l) | 1 \le a+2k \le n_1, 1 \le b+2l \le n_2\}. \tag{6.44}$$

For a clique $c \in \mathcal{C}$, one denotes by $V_c(\mathbf{y})$ the local potential, which is an arbitrary bounded function that depends only on the values of $\mathbf{y}$ in the clique c, $V_c(\mathbf{y}) = V_c(\mathbf{y}_c)$. The PI reminds that V_c may also depend on $\mathbf{x}$ in an arbitrary fashion. It is time now to introduce a local form of the distortion function as

$$D(\mathbf{y}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{y}_c). \tag{6.45}$$

The important fact is that D is a sum of functions with a small support. Let us express the local characteristics (6.24) in terms of this newly-defined form (6.45):

$$\pi_\lambda(Y_i = y_i' | \mathbf{y}_{\sim i}) = \frac{\exp(-\lambda \sum_{c \in \mathcal{C}} V_c(y_i' \mathbf{y}_{\sim i}))}{\sum_{t_i \in \mathcal{I}_i} \exp(-\lambda \sum_{c \in \mathcal{C}} V_c(t_i \mathbf{y}_{\sim i}))} \tag{6.46}$$

$$\stackrel{(a)}{=} \frac{\exp(-\lambda \sum_{c \in \mathcal{C}(i)} V_c(y_i' \mathbf{y}_{\sim i}))}{\sum_{t_i \in \mathcal{I}_i} \exp(-\lambda \sum_{c \in \mathcal{C}(i)} V_c(t_i \mathbf{y}_{\sim i}))}, \tag{6.47}$$

where $\mathcal{C}(i) = \{c \in \mathcal{C} | i \in c\}$, $i = 1, \ldots, n$. Equality (a) holds because $V_c(t_i \mathbf{y}_{\sim i})$ does not depend on t_i for cliques $c \notin \mathcal{C}(i)$ as they do not contain the ith element. Thus, the terms V_c for such cliques cancel from (6.47). This has a profound impact on the local characteristics, making the realization of Y_i *independent* of changes made outside of the union of cliques containing pixel i and thus outside of the neighborhood $\eta(i)$. For the cross-neighborhood system from Example 13, changes made to pixels belonging to the sublattice $\mathcal{S}_e$ do not interact and thus the Gibbs sampler can be parallelized by first updating *all* pixels from this sublattice in parallel and then updating in parallel *all* pixels from $\mathcal{S}_o$.[22]

[22]The Gibbs random field described by the joint distribution $\pi_\lambda(\mathbf{y})$ with distortion (6.45) becomes a Markov random field on the same neighborhood system. This follows from the Hammersley-Clifford theorem [92].

Algorithm 2 One sweep of a Gibbs sampler for embedding m-bit message (payload-limited sender).

Require: $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_s$ {mutually disjoint sublattices}
1: **for** $k = 1$ to s **do**
2: **for** every $i \in \mathcal{S}_k$ **do**
3: Use (6.48) to calculate cost of changing $y_i \to y_i' \in \mathcal{I}_i$
4: **end for**
5: Embed m/s bits while minimizing $\sum_{i \in \mathcal{S}_k} \rho_i(y_i' \mathbf{y}_{\sim i})$.
6: Update $\mathbf{y}_{\mathcal{S}_k}$ with new values and keep $\mathbf{y}_{\sim \mathcal{S}_k}$ unchanged.
7: **end for**
8: **return** $\mathbf{y}$

The possibility to update all pixels in each sublattice all at once provides a recipe for constructing practical embedding schemes. Assume $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_s$ with mutually disjoint sublattices. The PI first describes the actions of a payload-limited sender (follow the pseudo-code in Algorithm 2).

6.5.1. *Payload-limited sender.* The sender divides the payload of m bits into s equal parts of m/s bits, computes the local distortions

$$\rho_i(y_i' \mathbf{y}_{\sim i}) = \sum_{c \in \mathcal{C}(i)} V_c(y_i' \mathbf{y}_{\sim i}) \tag{6.48}$$

for pixels $i \in \mathcal{S}_1$, and embeds the first message part in $\mathcal{S}_1$. Then, it updates the local distortions of all pixels from $\mathcal{S}_2$ and embeds the second part in $\mathcal{S}_2$, updates the local distortions again, embeds the next part in $\mathcal{S}_3$, etc. Because the embedding changes in each sublattice do not interact, the embedding can be realized as discussed in Section 6.3. After all sublattices are processed it will mean that one embedding sweep was completed. By repeating these embedding sweeps,[23] the resulting modified images will converge to a sample from π_λ.

The embedding in sublattice $\mathcal{S}_k$ will introduce embedding changes with probabilities (6.15), where the value of λ_k is determined by the individual distortions $\{\rho_i(y_i' \mathbf{y}_{\sim i}) | i \in \mathcal{S}_k\}$ (6.48) to satisfy the payload constraint of embedding m/s bits in the kth sublattice (again, e.g., using a binary search for λ_k). Because each sublattice extends over a different portion of the cover image while splitting the payload evenly across the sublattices, λ_k may slightly vary with k because of variations in the individual distortions. This represents a deviation from the Gibbs sampler. Fortunately, the sublattices can often be chosen so that the image does not differ too much on every sublattice, which will guarantee that the sets of individual distortions $\{\rho_i(y_i' \mathbf{y}_{\sim i}) | i \in \mathcal{S}_k\}$ are also similar across the sublattices. Thus, with an increased number of sweeps, λ_k will converge to an approximately common value and the whole process represents a correct version of the Gibbs sampler.

In binary embedding ($\mathcal{I}_i = \{x_i^{(0)}, x_i^{(1)}\}$), note that the two distortions $\rho_i^{(0)}(x_i^{(0)} \mathbf{y}_{\sim i}) = D(x_i^{(0)} \mathbf{y}_{\eta(i)})$, $\rho_i^{(1)}(x_i^{(1)} \mathbf{y}_{\sim i}) = D(x_i^{(1)} \mathbf{y}_{\eta(i)})$ at pixel i depend on the current pixel values in its neighborhood $\eta(i)$. Therefore, both $\rho_i^{(0)}$ and $\rho_i^{(1)}$ can be non-zero at the same time and one can even have $\rho_i^{(1)} < \rho_i^{(0)}$. It is the neighborhood of i that ultimately determines whether or not it is beneficial to preserve the value of the pixel!

6.5.2. *Distortion-limited sender.* A similar approach can be used to implement the distortion-limited sender with a distortion limit D_ϵ. Consider a simulation of such embedding by a Gibbs sampler with the correct λ (obtained from a binary search as described in Section 6.4.2) on the sublattice $\mathcal{S}_k \subset \mathcal{S}$. Assuming again that all sublattices have the same distortion properties, the distortion obtained from cliques containing pixels from $\mathcal{S}_k$ should be proportional to the number of such cliques. Formally,

$$E_{\pi_\lambda(\mathbf{Y}_{\mathcal{S}_k} | \mathbf{Y}_{\sim \mathcal{S}_k})}[D] = D_\epsilon \frac{|\{c \in C | c \cap \mathcal{S}_k \neq \emptyset\}|}{|C|}. \tag{6.49}$$

As described in Algorithm 3, the sender can realize this by embedding as many bits to every sublattice as possible while achieving the distortion (6.49). Note that one does not need to compute the partition function for every image in order to realize the embedding. Moreover, in practice when the embedding is implemented using syndrome-trellis codes [26], the search for the correct parameter λ, as described in Section 6.4.2, is not needed either as long as the distortion properties of every sublattice are the same. This is because the codes need the local *distortion* $\rho_i(y_i' \mathbf{y}_{\sim i})$ (6.48) at each lattice pixel i and not the embedding probabilities. (This eliminates the need for λ.)

[23] After each embedding sweep, at each pixel the previous change is *erased* and the pixel is reconsidered again, just like in the Gibbs sampler.

Algorithm 3 One sweep of a Gibbs sampler for a distortion-limit sender, $E_{\pi_\lambda}[D] = D_\epsilon$.

Require: $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_s$ {mutually disjoint sublattices}
1: **for** $k = 1$ to s **do**
2: **for** every $i \in \mathcal{S}_k$ **do**
3: Use (6.48) to calculate cost of changing $y_i \to y_i' \in \mathcal{I}_i$
4: **end for**
5: Embed m_k bits while $\sum_i \rho_i(y_i' \mathbf{y}_{\sim i}) = D_\epsilon \times |\{c \in \mathcal{C} | c \cap \mathcal{S}_k \neq \emptyset\}|/|\mathcal{C}|$.
6: Update $\mathbf{y}_{\mathcal{S}_k}$ with new values and keep $\mathbf{y}_{\sim \mathcal{S}_k}$ unchanged.
7: **end for**
8: **return** $\mathbf{y}$ and $\sum_k m_k$ {stego image and number of bits}

The issue of the minimal sufficient number of embedding sweeps for both algorithms needs to be studied specifically for each distortion measure (see the discussion in the experimental Section 6.7). By replacing a specific practical embedding method with a simulator of optimal embedding, one can simulate the impact of optimal algorithms (for both senders) without having to determine the value of the parameter λ as described in Section 6.4.2. It is still necessary to compute λ_k for each sublattice $\mathcal{S}_k$ to obtain the probabilities of modifying each pixel (6.15), but this can be done as described in Section 6.3 without having to use the Gibbs sampler or the thermodynamic integration.

Finally, the PI comments on how to handle wet pixels within this framework. Since it is assumed that the distortion is bounded ($|D(\mathbf{y})| < K$ for all $\mathbf{y} \in \mathcal{Y}$), wet pixels are handled by forcing $\mathcal{I}_i = \{x_i\}$. Because this knowledge may not be available to the decoder in practice, practical coding schemes should treat them either by setting $\rho_i(y_i) = \infty$ or to some large constant for $y_i \neq x_i$ (for details, see [26]).

6.5.3. *Practical limits of the Gibbs sampler.* Thanks to the bounds established in Section 6.1, it is now known that the maximal payload that can be embedded in this manner is the entropy of π_λ (6.11). Assuming the embedding proceeds on the bound for the individual sublattices, the question is how close the total payload embedded in the image is to $H(\pi_\lambda)$. Following the Gibbs sampler, the configuration of the stego image will converge to a sample $\mathbf{y}$ from π_λ. Let us now go through one more sweep with $\mathbf{y}^{[k]}$ denoting the stego image before starting embedding in sublattice $\mathcal{S}_k$, $k = 1, \ldots, s$. In each sublattice, the following payload is embedded:

$$H\big(\mathbf{Y}_{\mathcal{S}_k} \big| \mathbf{Y}_{\sim \mathcal{S}_k} = \mathbf{y}^{[k]}_{\sim \mathcal{S}_k}\big). \tag{6.50}$$

The following result from information theory is now used: For any random variables $X_1, \ldots, X_s$,

$$\sum_{k=1}^{s} H(X_k | X_{\sim k}) \leq H(X_1, \ldots, X_s), \tag{6.51}$$

with equality only when all variables are independent.[24] Thus, in general

$$H^-(\mathbf{Y}) \triangleq \sum_{k=1}^{s} H\big(\mathbf{Y}_{\mathcal{S}_k} \big| \mathbf{Y}_{\sim \mathcal{S}_k} = \mathbf{y}^{[k]}_{\sim \mathcal{S}_k}\big) < H(\mathbf{Y}) = H(\pi_\lambda). \tag{6.52}$$

The term $H^-(\mathbf{Y})$ is recognized as the erasure entropy [84, 85] and it is equal to the conditional entropy $H(\mathbf{Y}^{(l+1)}|\mathbf{Y}^{(l)})$ (entropy rate) of the Markov process defined by our Gibbs sampler (c.f., (6.35)), where $\mathbf{Y}^{(l)}$ is the random variable obtained after l sweeps of the Gibbs sampler.

The erasure-entropy inequality (6.52) means that the embedding scheme will be suboptimal, unable to embed the maximal payload $H(\pi_\lambda)$. The actual loss can be assessed by evaluating the entropy of $H(\pi_\lambda)$, e.g., using the algorithms described in Section 6.4. An example of such comparison is presented in Section 6.7.3.

The last remaining issue is the choice of the potentials V_c. In the next section, the PI shows one example where V_c are chosen to tie the principle of minimal embedding distortion to the preservation of the cover-source model. The PI also describes a specific embedding method and subjects it to experiments using blind steganalyzers.

[24] For $k = 2$, this result follows immediately from $H(X_1|X_2) + H(X_2|X_1) = H(X_1, X_2) - I(X_1; X_2)$. The result for $s > 2$ can be obtained by induction over s.

6.6. **Practical embedding constructions.** Here, a practical embedding method that uses the theory developed so far is described. First and foremost, the potentials V_c should measure the detectability of embedding changes. There is substantial freedom in choosing them and the design may utilize reasoning based on theoretical cover source models as well as heuristics stemming from experiments using blind steganalyzers. The proper design of potentials is a complicated subject in itself and is beyond the scope of this effort, whose main purpose was to introduce a general framework rather than optimizing the design. Here, the PI describes a specific example of a more general approach that builds upon the latest results in steganography and steganalysis and one that provided an opportunity to validate the proposed framework by showing an improvement over the current state of the art in Section 6.7.

6.6.1. *Additive approximation.* As argued in the introduction, the steganography design principles based on model preservation and on minimizing distortion coincide when the distortion is defined as a norm of the difference of feature vectors used to model cover images:

$$D(\mathbf{y}) = ||f(\mathbf{x}) - f(\mathbf{y})|| \triangleq \sum_{k=1}^{d} w_k |f_k(\mathbf{x}) - f_k(\mathbf{y})|. \tag{6.53}$$

Here, $f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_d(\mathbf{x})) \in \mathbb{R}^d$ is a d-dimensional feature vector of image $\mathbf{x}$ and $\mathbf{w} = (w_1, \ldots, w_d)$ are weights. The properties of D defined in this manner depend on the properties of the functions f_k. In general, however, D is not additive. In the past, steganographers were forced to use some *additive approximation* of D to realize the embedding in practice. A general method for turning an arbitrary distortion measure into an additive proceeds is:

$$\hat{D}(\mathbf{y}) = \sum_{i=1}^{n} D(y_i \mathbf{x}_{\sim i}). \tag{6.54}$$

Embedding with the additive measure $\hat{D}$ can be simulated (and realized) as explained in Section 6.3. The approximation, of course, ensues a capacity loss due to a mismatch in the minimized distortion function. Thanks to the methods introduced in Section 6.4.2, this loss can now be contrasted against the rate–distortion bound for the original measure D. However, one cannot build a practical scheme unless D can be written as a sum of *local* potentials. Next, the PI explains how to turn D into this form using the idea of a bounding distortion.

6.6.2. *Bounding distortion.* Most features used in steganalysis can be written as a sum of locally-supported functions across the image

$$f_k(\mathbf{x}) = \sum_{c \in \mathcal{C}} f_c^{(k)}(\mathbf{x}), \quad k = 1, \ldots, d. \tag{6.55}$$

For example, the kth histogram bin of image $\mathbf{x}$ can be written using the Iverson bracket as

$$h_k(\mathbf{x}) = \sum_{i \in \mathcal{S}} [x_i = k], \tag{6.56}$$

while the klth element of a horizontal co-occurrence matrix

$$C_{k,l}(\mathbf{x}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2 - 1} [x_{i,j} = k][x_{i,j+1} = l] \tag{6.57}$$

is a sum over horizontally adjacent pixels (horizontal two-pixel cliques). For such locally-supported features, one can obtain an upper bound on $D(\mathbf{y}) = ||f(\mathbf{x}) - f(\mathbf{y})||$, $\mathbf{y} \in \mathcal{Y}$, that has the required form:

$$||f(\mathbf{x}) - f(\mathbf{y})|| = \sum_{k=1}^{d} w_k \left| \sum_{c \in \mathcal{C}} f_c^{(k)}(\mathbf{x}) - \sum_{c} f_c^{(k)}(\mathbf{y}) \right| \tag{6.58}$$

$$\leq \sum_{k=1}^{d} w_k \sum_{c \in \mathcal{C}} |f_c^{(k)}(\mathbf{x}) - f_c^{(k)}(\mathbf{y})| \tag{6.59}$$

$$= \sum_{c \in \mathcal{C}} \sum_{k=1}^{d} w_k |f_c^{(k)}(\mathbf{x}) - f_c^{(k)}(\mathbf{y})| \tag{6.60}$$

$$= \sum_{c \in \mathcal{C}} V_c(\mathbf{y}), \tag{6.61}$$

where

$$V_c(\mathbf{y}) = \sum_{k=1}^{d} w_k |f_c^{(k)}(\mathbf{x}) - f_c^{(k)}(\mathbf{y})|. \tag{6.62}$$

Following our convention explained in Section 6.1, the PI describes the methodology for a fixed cover image $\mathbf{x}$ and thus does not make the dependence of V_c on $\mathbf{x}$ explicit. The sum $\sum_{c\in\mathcal{C}} V_c(\mathbf{y})$ will be called the *bounding distortion.*

The PI now provides a specific example of this approach. The choice is motivated by the desire to work with a modern, well-established feature set so that later, in Section 6.7, one can validate the usefulness of the proposed framework by constructing a high-capacity steganographic method undetectable using current state-of-the-art steganalyzer. The motivation and justification of the feature set appears in [66]. It is a slight modification of the SPAM set [64], which is the basis of the current most reliable blind steganalyzer in the spatial domain. The features are constructed by considering the differences between neighboring pixels (e.g., horizontally adjacent pixels) as a higher-order Markov chain and taking the sample joint probability matrix (co-occurrence matrix) as the feature. The advantage of using the joint matrix instead of the transition probability matrix is that the norm of the feature difference can be readily upper-bounded by the desired local form (6.62).

To formally define the feature for an $n_1 \times n_2$ image $\mathbf{x}$, let us consider the following co-occurrence matrix computed from horizontal pixel differences $D_{i,j}^{\rightarrow}(\mathbf{x}) = x_{i,j+1} - x_{i,j}$, $i = 1,\dots,n_1, j = 1,\dots,n_2 - 1$:

$$A_{k,l}^{\rightarrow}(\mathbf{x}) = \frac{1}{n_1(n_2-2)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2-2} [(D_{i,j}^{\rightarrow}, D_{i,j+1}^{\rightarrow})(\mathbf{x}) = (k,l)]. \tag{6.63}$$

For compactness, in (6.63) the argument of the Iverson bracket is abbreviated from $D_{i,j}^{\rightarrow}(\mathbf{x}) = k \,\&\, D_{i,j+1}^{\rightarrow}(\mathbf{x}) = l$ to $(D_{i,j}^{\rightarrow}, D_{i,j+1}^{\rightarrow})(\mathbf{x}) = (k,l)$. Clearly, $A_{i,j}^{\rightarrow}(\mathbf{x})$ is the normalized count of neighboring triples of pixels $\{x_{i,j}, x_{i,j+1}, x_{i,j+2}\}$ with differences $x_{i,j+1} - x_{i,j} = k$ and $x_{i,j+2} - x_{i,j+1} = l$ in the entire image. The superscript arrow "$\rightarrow$" denotes the fact that the differences are computed by subtracting the left pixel from the right one. Similarly,

$$A_{k,l}^{\leftarrow}(\mathbf{x}) = \frac{1}{n_1(n_2-2)} \sum_{i=1}^{n_1} \sum_{j=3}^{n_2} [(D_{i,j}^{\leftarrow}, D_{i,j-1}^{\leftarrow})(\mathbf{x}) = (k,l)] \tag{6.64}$$

with $D_{i,j}^{\leftarrow}(\mathbf{x}) = x_{i,j-1} - x_{i,j}$. By analogy, one can define vertical, diagonal, and minor diagonal matrices $A_{k,l}^{\downarrow}$, $A_{k,l}^{\uparrow}$, $A_{k,l}^{\nearrow}$, $A_{k,l}^{\swarrow}$, $A_{k,l}^{\searrow}$, $A_{k,l}^{\nwarrow}$. All eight matrices are sample joint probabilities of observing the differences k and l between three consecutive pixels along a certain direction. Due to the antisymmetry $D_{i,j}^{\rightarrow}(\mathbf{x}) = -D_{i,j+1}^{\leftarrow}(\mathbf{x})$ only $A_{k,l}^{\rightarrow}$, $A_{k,l}^{\nearrow}$, $A_{k,l}^{\uparrow}$, $A_{k,l}^{\nwarrow}$ are needed since $A_{k,l}^{\rightarrow} = A_{-l,-k}^{\leftarrow}$, and similarly for other matrices.

Because neighboring pixels in natural images are strongly dependent, each matrix exhibits a sharp peak around $(k,l) = (0,0)$ and then quickly falls off with increasing k and l. When such matrices are used for steganalysis [64], they are truncated to a small range, such as $-T \le k,l \le T$, $T = 4$, to prevent the onset of the "curse of dimensionality." On the other hand, in steganography one can use large-dimensional models ($T = 255$) because it is easier to preserve a model than to learn it.[25] Another reason for using a high-dimensional feature space is to avoid "overtraining" the embedding algorithm to a low-dimensional model as such algorithms may become detectable by a slightly modified feature set, an effect already reported in the DCT domain [56].

[25]Similar reasoning for constructing the distortion function was used in the HUGO algorithm [66].

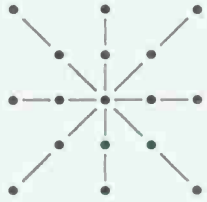

FIGURE 6.5. The union of all 12 cliques consisting of three pixels arranged in a straight line in the 5×5 square neighborhood.

By embedding a message, $A^{\rightarrow}_{k,l}(\mathbf{x})$ is modified to $A^{\rightarrow}_{k,l}(\mathbf{y})$. The differences between the features will thus serve as a measure of embedding impact closely tied to the model (the indices i and j run from 1 to n_1 and $n_2 - 2$, respectively):

$$|A^{\rightarrow}_{k,l}(\mathbf{y}) - A^{\rightarrow}_{k,l}(\mathbf{x})| = \tag{6.65}$$

$$= \frac{1}{n_1(n_2-2)}\Big|\sum_{i,j}[(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1})(\mathbf{y}) = (k,l)] \tag{6.66}$$

$$- [(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1})(\mathbf{x}) = (k,l)]\Big| \tag{6.67}$$

$$\leq \frac{1}{n_1(n_2-2)}\sum_{i,j}\Big|[(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1})(\mathbf{y}) = (k,l)] \tag{6.68}$$

$$- [(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1}(\mathbf{x}) = (k,l)]\Big| \tag{6.69}$$

$$= \sum_{c\in\mathcal{C}^{\rightarrow}} H_c^{(k,l)\rightarrow}(\mathbf{y}), \tag{6.70}$$

where the PI defined the following locally-supported functions

$$H_c^{(k,l)\rightarrow}(\mathbf{y}) = \frac{1}{n_1(n_2-2)}$$

$$\cdot\Big|[(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1})(\mathbf{y}) = (k,l)] - [(D^{\rightarrow}_{i,j}, D^{\rightarrow}_{i,j+1})(\mathbf{x}) = (k,l)]\Big| \tag{6.71}$$

on all horizontal cliques $\mathcal{C}^{\rightarrow} = \{c | c = \{(i,j),(i,j+1),(i,j+2)\}\}$. Notice that the absolute value had to be pulled into the sum to give the potentials a small support.

Since the other three matrices can be written in this manner as well, one can write the distortion function in the following final form

$$D(\mathbf{y}) = \sum_{c\in\mathcal{C}} V_c(\mathbf{y}), \tag{6.72}$$

now with $\mathcal{C} = \mathcal{C}^{\rightarrow} \cup \mathcal{C}^{\nearrow} \cup \mathcal{C}^{\uparrow} \cup \mathcal{C}^{\nwarrow}$, the set of three-pixel cliques along all four directions, and

$$V_c(\mathbf{y}) = \sum_{k,l} w_{k,l} H_c^{(k,l)\rightarrow}(\mathbf{y}), \text{ for each clique } c \in \mathcal{C}^{\rightarrow}, \tag{6.73}$$

and similarly for the other three clique types. Notice that there are again weights $w_{k,l} > 0$ in the definition of V_c that can be adjusted according to how sensitive steganalysis is to the individual differences. For example, if a certain difference pair (k,l) varies significantly over cover images, by assigning it a smaller weight one can allow it to be modified more often, while those differences that are stable across covers but sensitive to embedding should be intuitively assigned a larger value so that the embedding does not modify them too much.

To complete the picture, the neighborhood system here is formed by 5×5 neighborhoods and thus the index set can be decomposed into nine disjoint sublattices $\mathcal{S} = \bigcup_{ab} \mathcal{S}_{ab}$, $1 \leq a, b \leq 3$,

$$\mathcal{S}_{ab} = \{(a+3k, b+3l) | 1 \leq a+3k \leq n_1, 1 \leq b+3l \leq n_2\}. \tag{6.74}$$

To better explain the effect of embedding changes on the distortion, realize that each pixel belongs to three horizontal, three vertical, three diagonal, and three minor-diagonal cliques. When a single pixel $x_{i,j}$ is changed, it affects only the 12 potentials whose clique contains $x_{i,j}$. For example if the original pixel values $c_0 = \{x_{i,j}, x_{i,j+1}, x_{i,j+2}\}$ had differences k, l, and the pixel value changed from $x_{i,j}$ to $y_{i,j} = x_{i,j} + 1$. Then, the pixel differences will be modified to $k-1, l$. Considering

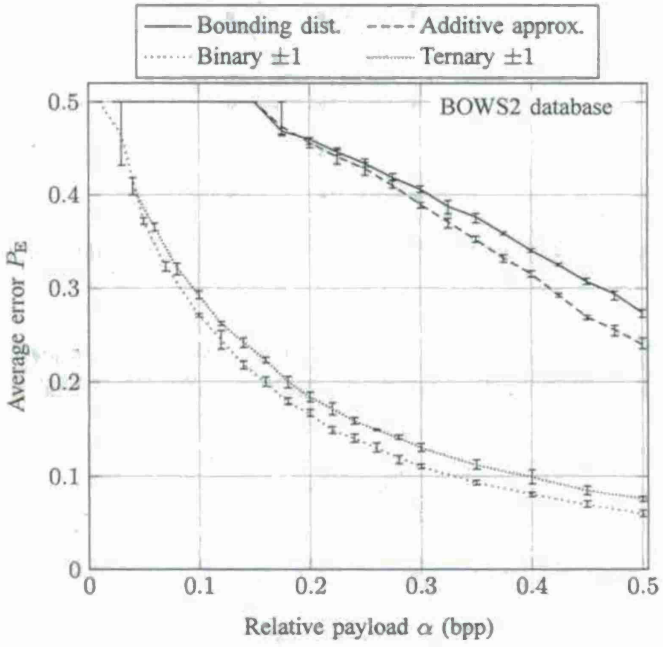

FIGURE 6.6. Comparison of ±1 embedding with optimal binary and ternary coding with binary embedding algorithms based on the Gibbs construction with a bounding distortion and the additive approximation as described in Section 6.7.1. The error bars depict the minimum and maximum steganalyzer error P_E over five runs of SVM classifiers with different division of images into training and testing set.

just the contribution from $H_{c_0}^{(k,l)\rightarrow}$ to the potential V_{c_0} (6.73), it will increase by the sum of $w_{k,l}$ (the pair k,l is leaving cover) and $w_{k-1,l}$ (a new pair appears in the stego image).

6.6.3. *Other options.* The framework presented in this report allows the sender to formulate the local potentials directly instead of obtaining them as the bounding distortion. For example, the cliques and their potentials may be determined by the local image content or by learning the cover source using the method of fields of experts [70]. The merit of these possibilities can be evaluated by steganalyzers trained on a large set of images. The important question of optimizing the local potential functions w.r.t. statistical detectability is an important direction the PI intends to explore in the future.

6.7. **Experiments.** In this section, the PI validates the proposed framework experimentally and includes a comparison between simple steganographic algorithms, such as binary and ternary ±1 embedding and steganography implemented via the bounding distortion and the additive approximation (6.54). For the case of the bounding distortion, the capacity loss w.r.t. the optimal payload given by $H(\pi_\lambda)$ is evaluated by means of the thermodynamic integration algorithm from Section 6.4.2.

6.7.1. *Tested embedding methods.* For the methods based on additive approximation and the bounding distortion, the PI used as a feature vector the joint probability matrix $A_{k,l,m}^{\rightarrow}(\mathbf{x})$ defined similarly as in (6.63) with the difference vector computed from *four* consecutive pixels $(D_{i,j}^{\rightarrow}, D_{i,j+1}^{\rightarrow}, D_{i,j+2}^{\rightarrow}) = (k,l,m)$. As above, four such matrices corresponding to four spatial directions were computed. The matrices were used at their full size $T = 255$ leading to model dimensionality $d = 4 \times 511^3 \approx 5 \cdot 10^8$.

The weights were chosen to be small for those triples $(D_{i,j}^{\rightarrow}, D_{i,j+1}^{\rightarrow}, D_{i,j+2}^{\rightarrow}) = (k,l,m)$ that occur infrequently in images and large for frequented triples. Following the recommendation described in [66], since the frequency of occurrence of the triples falls off quickly with their norm, the weights are chosen as

$$w_{k,l,m} = \left(\sigma + \sqrt{k^2 + l^2 + m^2}\right)^{-\theta}, \tag{6.75}$$

with $\theta = 1$ and $\sigma = 1$. The purpose of the weights is to force the embedding algorithm to modify those parts of the model that are difficult to model accurately, forcing thus the steganalyst to use a more accurate model. Here, the advantage goes to the steganographer, because preserving a high-dimensional feature vector is more feasible than accurately modeling it.

Because the neighborhood $\eta(i)$ in this case contains 7×7 pixels, the image was divided into 16 square sublattices on which embedding was carried out independently. The PI tested binary embedding, $\mathcal{I}_i = \{x_i, x_i'\}$, where x_i' was selected randomly and uniformly from $\{x_i - 1, x_i + 1\}$ and then fixed for all experiments with cover $\mathbf{x}$. The payload-limited sender was simulated using the Gibbs sampler constrained to only two sweeps. Increasing the number of sweeps did not lead to further improvement. The curiously low number of sweeps sufficient to properly implement the Gibbs sampler is most likely due to the fact that the dependencies dictated by the bounding distortion are rather weak. The simulation of embedding for one image took less than 5 seconds when implemented in C++ on a single-processor PC.

To summarize, the following four steganographic methods were tested:

(1) Binary embedding using the Gibbs construction with sets $\mathcal{I}_i = \{x_i, x_i'\}$ and bounding distortion (6.72) of (6.53) with weights (6.75) for the $d = 4 \times 511^3$-dimensional feature space given by matrices $A^{\rightarrow}_{k,l,m}$, $A^{\nearrow}_{k,l,m}$, $A^{\uparrow}_{k,l,m}$, $A^{\nwarrow}_{k,l,m}$.
(2) Additive approximation (6.54) of (6.53) for the same sets $\mathcal{I}_i$, feature space, and norm as in 1).
(3) Binary ± 1 embedding with the same sets $\mathcal{I}_i$ equipped with a matrix embedding scheme operating on the binary bound.
(4) Ternary ± 1 embedding with $\mathcal{I}_i = \{x_i - 1, x_i, x_i + 1\}$ equipped with a ternary matrix embedding scheme operating on the ternary bound (the bounds appear, e.g., in [29]).

Practical near-optimal codes for the two ± 1 embedding methods can be found in [31] and [98].

6.7.2. *Testing methodology and final results.* Following the separation principle, the PI now studies the security of all schemes when operating on the rate–distortion bound. All tests were carried out on the BOWS2 database [2] containing approximately 10800 grayscale images with a fixed size of 512×512 pixels coming from rescaled and cropped natural images of various sizes. Steganalysis was implemented using the second-order SPAM feature set with $T = 3$ [64]. The image database was evenly divided into a training and a testing set of cover and stego images, respectively. A soft-margin support-vector machine was trained using the Gaussian kernel. The kernel width and the penalty parameter were determined using five-fold cross validation on the grid $(C, \gamma) \in \{(10^k, 2^j) | k \in \{-3, \dots, 4\}, j \in \{-L - 3, \dots, -L + 3\}\}$, where $L = \log_2 d$ is the binary logarithm of the number of features.

The results are reported using the minimum average classification error P_{E}. Smaller values of P_{E} correspond to better steganalysis and thus larger statistical detectability (lower security).

Figure 6.6 displays the comparison of all four embedding methods listed above. The methods based on the the bounding distortion and the additive approximation (denoted as "Bounding dist." and "Additive approx.") are completely undetectable for payloads smaller than 0.15 bpp, which suggests that the embedding changes are made in pixels not covered by the SPAM features. Since both schemes are binary with $\mathcal{I}_i = \{x_i, x_i'\}$ with x_i' randomly chosen from $\{x_i - 1, x_i + 1\}$, they become equivalent to simple binary ± 1 embedding (Method 3) as $\alpha \to 1$ and thus become detectable. Comparing the capacity, both schemes allow communicating ten times larger payloads with $P_{\text{E}} = 40\%$ as compared to ternary ± 1 embedding. The advantage of using the Gibbs sampler with the bounding distortion over the additive approximation becomes more evident for larger payloads, where the embedding changes start to interact. This confirms the expectation that in this range the additive approximation is unable to cope with the interactions among changes and thus its detectability increases. This result, however, may change for different distortion measures and cover sources. The fact that the Gibbs sampler with bounding distortion did not bring a substantial performance improvement over the additive approximation indicates that the interactions among embedding changes are in general quite weak (at least as far as they are captured by the bounding distortion). The low strength of interactions also explains why only two sweeps of the Gibbs sampler were sufficient in practice.

6.7.3. *Analysis of upper bounds.* As described in Section 6.5.3, Algorithm 2 for the payload-limited sender is unable to embed the optimal payload of $H(\pi_\lambda)$ for three reasons. The performance may be affected by the small number of sweeps of the Gibbs sampler, the parameter λ may vary slightly among the sublattices, and the algorithm embeds the erasure entropy $H^-(\pi_\lambda) \leq H(\pi_\lambda)$. The combined effect of these factors is of great importance for practitioners and is evaluated below for two images using the Gibbs sampler and the thermodynamic integration as explained in Section 6.4.2.

Since the Gibbs construction depends on the cover image $\mathbf{x}$, the PI present the results for two grayscale images of size 512×512 pixels coming from two different sources. The test image "0.png" is from the BOWS2 database and "Lenna" was obtained from `http://en.wikipedia.org/wiki/File:Lenna.png` and converted to grayscale using GNU Image Manipulation Program (GIMP). In both cases, the PI used the same sets $\mathcal{I}_i$ and the same feature set as in the previous section with the bounding distortion with weight parameters $\sigma = 1$ and $\theta = 1$.

The image "0.png" contains more areas with edges and textures than "Lenna" and thus for small distortions, it offers a larger capacity than "Lenna" because the weights (6.75) around edges and complex texture are small. This is apparent from the slopes of the rate–distortion bounds in Figure 6.7.

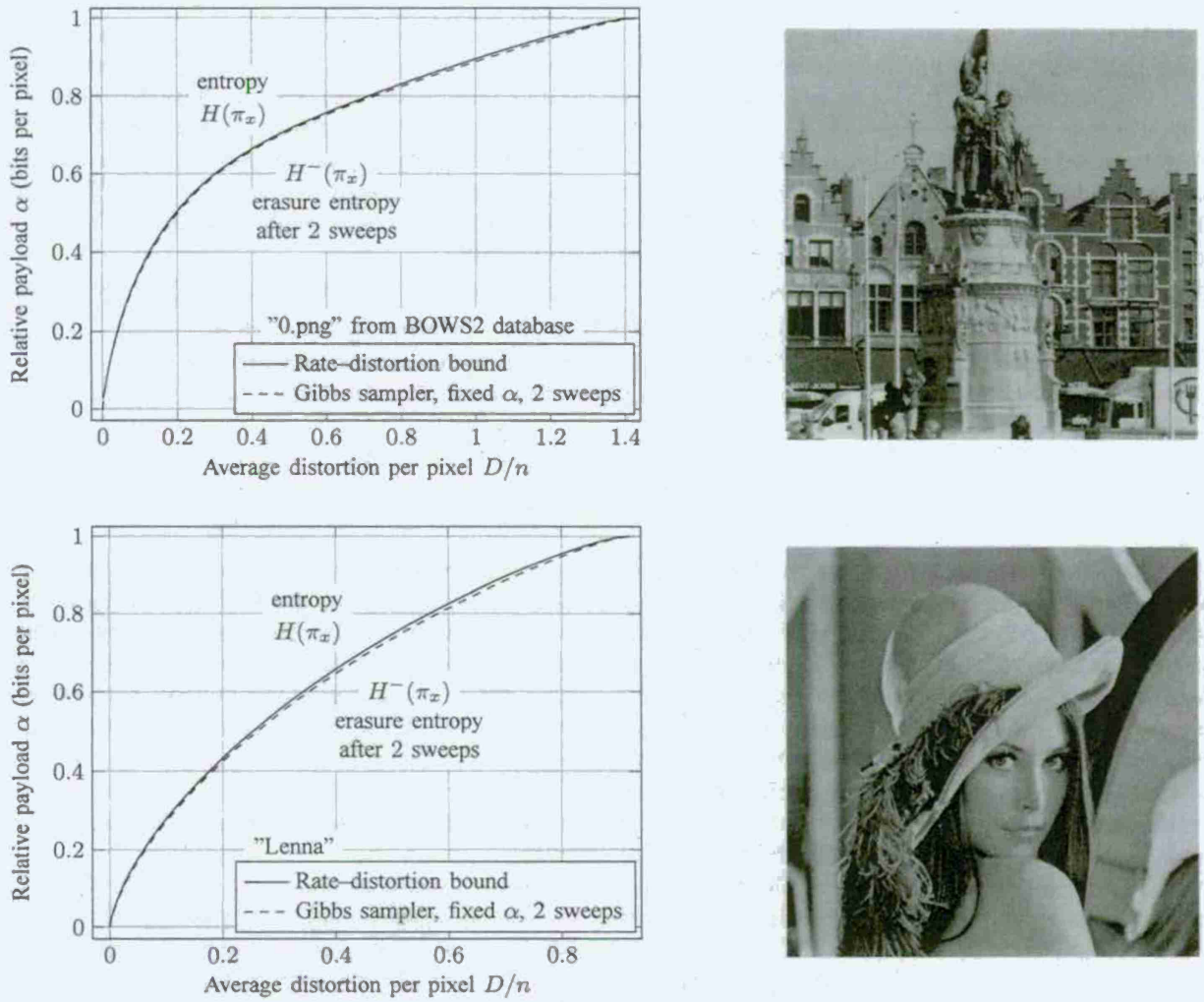

FIGURE 6.7. Comparison of the payload loss of Algorithm 2 for cover images "0.png" and "Lenna" shown on the right. The rate–distortion bounds were obtained using the Gibbs sampler (6.38) and the thermodynamic integration (6.40).

The same figure compares the rate–distortion performance of the payload-limited sender simulated by the Gibbs sampler with only two sweeps as described in Algorithm 2. For a given payload, the distortion was obtained as an average over 100 random messages. The comparison shows that the payload loss of Algorithm 2 to the optimal $H(\pi_\lambda)$ is quite small. Note that the erasure entropy, $H^-(\pi_\lambda)$, plotted in the figure has been computed over the sublattices after two sweeps and thus already contains the impact of all three factors discussed at the beginning of this section.

6.8. **Discussion.** Currently, the most successful principle for designing practical steganographic systems that embed in empirical covers is based on minimizing a suitably defined distortion measure. Implementation difficulties and a lack of practical embedding methods have so far limited the application of this principle to a rather special class of distortion measures that are additive over pixels. With the development of near-optimal low-complexity coding schemes, such as the syndrome-trellis codes [26], this direction has essentially reached its limits. It is a firm belief of the PI that further substantial increase in secure payload is possible only when the sender uses adaptive schemes that place embedding changes based on the local content, that dare to modify pixels in some regions by more than 1, and that consider interactions among embedding changes while preserving higher-order statistics among pixels. This section describes a contribution which is an important step in this direction.

It offers the steganographer a complete methodology for embedding while minimizing an arbitrarily defined distortion measure D. The absence of any restrictions on D means that the remaining task left to the sender is to find a distortion measure that correlates with statistical detectability. An appealing possibility is to define D as a weighted norm of

the difference between cover and stego feature vectors used in steganalysis. This immediately connects the principle of minimum-distortion steganography with the concept of model preservation which has so far been limited to low-dimensional models. Being able to preserve a large-dimensional model gives the steganographer a great advantage over the steganalyst because of the difficulties associated with learning a high-dimensional cover source model using statistical learning tools.

The proposed framework is called the Gibbs construction and it connects steganography with statistical physics, which contributed with many practical algorithms. In particular, the Gibbs sampler combined with the thermodynamic integration can be used to derive the rate–distortion bound, simulate the impact of optimal embedding, and realize near-optimal embedding algorithms. These three tasks can be addressed separately (the so-called "separation principle") giving the sender a great amount of design flexibility as well as control over losses of practical schemes.

An important case elaborated in this section corresponds to D defined as a sum of local potentials over small pixel neighborhoods. Here, the optimal distribution of embedding modifications reduces to a Markov random field and the Gibbs sampler can be turned into a practical embedding algorithm able to consider dependencies among embedding changes. When D cannot be written as a sum of local potentials, practical (suboptimal) methods can be realized by approximating D either with an additive distortion measure or with local potentials. The problem of finding the best approximation for a given non-local D is of its own interest. The PI did not cover the task of minimizing the statistical detectability with respect to the distortion function completely due to its inherent complexity; it is left as part of future effort.

The proposed methodology was described both for a payload-limited sender and the distortion-limited sender. The former embeds a fixed payload in every image with minimal distortion, while the latter embeds the maximal payload for a given distortion in every image. The distortion-limited sender better corresponds to the intuition that, for a fixed statistical detectability, more textured or noisy images can carry a larger secure payload than smoother or simpler images. The fact that the size of the hidden message is driven by the cover image essentially represents a more realistic case of the batch steganography paradigm [50].

Note that the distortion measure is used only by the sender and thus does not need to be shared. The only information needed by the receiver to decode the message is its size which can be communicated separately in the same cover image. This opens up the intriguing possibility to develop embedding schemes able to learn the proper distortion function while observing the impact of embedding on the cover source.

Finally, the proposed methodology can be applied to other data hiding problems where the statistical detectability constraint could be replaced by a perceptual distortion constraint. The implicit premise of this work is the direct relationship between the distortion function D and statistical detectability. Designing (and possibly learning) the distortion measure for a given cover source is an interesting problem by itself and was addressed by the PI in her last publication [24] developed under this effort and made possible by the No-Cost Extension. C++ implementation with Matlab wrappers of STCs and multi-layered STCs are available at `http://dde.binghamton.edu/download/syndrome/`.

References

[1] R. Anderson. Stretching the limits of steganography. In R. J. Anderson, editor, *Information Hiding, 1st International Workshop*, volume 1174 of *Lecture Notes in Computer Science*, pages 39–48, Cambridge, UK, May 30 – June 1, 1996. Springer-Verlag.

[2] P. Bas and T. Furon. BOWS-2. http://bows2.gipsa-lab.inpg.fr, July 2007.

[3] J. Bierbrauer. On Crandall's problem. Personal communication available from http://www.ws.binghamton.edu/fridrich/covcodes.pdf, 1998.

[4] J. Bierbrauer and J. Fridrich. Constructing good covering codes for applications in steganography. *LNCS Transactions on Data Hiding and Multimedia Security*, 4920:1–22, 2008.

[5] R. Böhme. Assessment of steganalytic methods using multiple regression models. In M. Barni, J. Herrera, S. Katzenbeisser, and F. Pérez-González, editors, *Information Hiding, 7th International Workshop*, Barcelona, Spain, June 6–8, 2005.

[6] R. Böhme. *Improved Statistical Steganalysis Using Models of Heterogeneous Cover Signals*. PhD thesis, Faculty of Computer Science, Technische Universität Dresden, Germany, 2008.

[7] R. Böhme and A. D. Ker. A two-factor error model for quantitative steganalysis. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, pages 59–74, San Jose, CA, January 16–19, 2006.

[8] R. Böhme and A. Westfeld. Breaking Cauchy model-based JPEG steganography with first order statistics. In P. Samarati, P. Y. A. Ryan, D. Gollmann, and R. Molva, editors, *Computer Security - ESORICS 2004. Proceedings 9th European Symposium on Research in Computer Security*, volume 3193 of Lecture Notes in Computer Science, pages 125–140, Sophia Antipolis, France, September 13–15, 2004. Springer, Berlin.

[9] C. Cachin. An information-theoretic model for steganography. In D. Aucsmith, editor, *Information Hiding, 2nd International Workshop*, volume 1525 of Lecture Notes in Computer Science, pages 306–318, Portland, OR, April 14–17, 1998. Springer-Verlag, New York.

[10] G. Cancelli and M. Barni. MPSteg-color: A new steganographic technique for color images. In T. Furon, F. Cayre, G. Doërr, and P. Bas, editors, *Information Hiding, 9th International Workshop*, volume 4567 of Lecture Notes in Computer Science, pages 1–15, Saint Malo, France, June 11–13, 2007. Springer-Verlag, Berlin.

[11] C. Chen and Y. Q. Shi. JPEG image steganalysis utilizing both intrablock and interblock correlations. In *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pages 3029–3032, May 2008.

[12] P. Comesaña and F. Pérez-Gonzáles. On the capacity of stegosystems. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 3–14, Dallas, TX, September 20–21, 2007.

[13] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.

[14] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, June 1996.

[15] R. Crandall. Some notes on steganography. *Steganography Mailing List*, available from `http://os.inf.tu-dresden.de/~westfeld/crandall.pdf`, 1998.

[16] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press, 2000.

[17] J. L. Doob. *Stochastic processes*. Wiley, New York, 1st edition, 1953.

[18] T. Filler. Important properties of normalized KL-divergence under HMC model. Technical report, DDE Lab, SUNY Binghamton, 2008. `http://dde.binghamton.edu/filler/kl-divergence-hmc.pdf`.

[19] T. Filler. Fisher information determines capacity of ε-secure steganography - proofs. Technical report, SUNY Binghamton, 2009. `http://dde.binghamton.edu/filler/pdf/Fill09ihwproofs.pdf`.

[20] T. Filler and J. Fridrich. Minimizing additive distortion functions with non-binary embedding operation in steganography.

[21] T. Filler and J. Fridrich. Binary quantization using belief propagation over factor graphs of LDGM codes. In *45th Annual Allerton Conference on Communication, Control, and Computing*, Allerton, IL, September 26–28, 2007.

[22] T. Filler and J. Fridrich. Wet ZZW construction for steganography. In *First IEEE International Workshop on Information Forensics and Security*, London, UK, December 6–9 2009.

[23] T. Filler and J. Fridrich. Gibbs construction in steganography. *IEEE Transactions on Information Forensics and Security*, 5(4):705–720, 2010.

[24] T. Filler and J. Fridrich. Design of adaptive steganographic schemes for digital images in spatial domain. volume 7880, 2011.

[25] T. Filler, J. Judas, and J. Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 2010. Under preparation.

[26] T. Filler, J. Judas, and J. Fridrich. Minimizing embedding impact in steganography using trellis-coded quantization. In N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XII*, volume 7541, pages 05–01–05–14, San Jose, CA, January 17–21, 2010.

[27] T. Filler, A. D. Ker, and J. Fridrich. The Square Root Law of steganographic capacity for Markov covers. In N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XI*, volume 7254, pages 08 1–08 11, San Jose, CA, January 18–21, 2009.

[28] J. Fridrich. Asymptotic behavior of the ZZW embedding construction. *IEEE Transactions on Information Forensics and Security*, 4(1):151–153, March 2009.

[29] J. Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.

[30] J. Fridrich, P. L. ek, and D. Soukal. On steganographic embedding efficiency. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of Lecture Notes in Computer Science, pages 282–296, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.

[31] J. Fridrich and T. Filler. Practical methods for minimizing embedding impact in steganography. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, pages 02–03, San Jose, CA, January 29–February 1, 2007.

[32] J. Fridrich and M. Goljan. Digital image steganography using stochastic modulation. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents V*, volume 5020, pages 191–202, Santa Clara, CA, January 21–24, 2003.

[33] J. Fridrich, M. Goljan, and D. Soukal. Perturbed quantization steganography. *ACM Multimedia System Journal*, 11(2):98–107, 2005.

[34] J. Fridrich, M. Goljan, and D. Soukal. Wet paper codes with improved embedding efficiency. *IEEE Transactions on Information Forensics and Security*, 1(1):102–110, 2006.

[35] J. Fridrich, M. Goljan, D. Soukal, and P. Lisoněk. Writing on wet paper. In T. Kalker and P. Moulin, editors, *IEEE Transactions on Signal Processing, Special Issue on Media Security*, volume 53, pages 3923–3935, October 2005. (journal version).

[36] J. Fridrich, M. Goljan, D. Soukal, and P. Lisoněk. Writing on wet paper. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 328–340, San Jose, CA, January 16–20, 2005.

[37] J. Fridrich, T. Pevný, and J. Kodovský. Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 3–14, Dallas, TX, September 20–21, 2007.

[38] J. Fridrich and D. Soukal. Matrix embedding for large payloads. *IEEE Transactions on Information Forensics and Security*, 1(3):390–394, 2006.

[39] S. I. Gel'fand and M. S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

[40] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[41] M. Goljan, J. Fridrich, and T. Holotyak. New blind steganalysis and its implications. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, pages 1–13, San Jose, CA, January 16–19, 2006.

[42] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[43] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2007. MPI Technical Report 157.

[44] I. Hen and N. Merhav. On the error exponent of trellis source coding. *IEEE Transactions on Information Theory*, 51(11):3734–3741, 2005.

[45] S. Hetzl and P. Mutzel. A graph-theoretic approach to steganography. In J. Dittmann, S. Katzenbeisser, and A. Uhl, editors, *Communications and Multimedia Security, 9th IFIP TC-6 TC-11 International Conference, CMS 2005*, volume 3677 of Lecture Notes in Computer Science, pages 119–128, Salzburg, Austria, September 19–21, 2005.

[46] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, volume II. Upper Saddle River, NJ: Prentice Hall, 1998.

[47] A. Ker and R. Böhme. Revisiting weighted stego-image steganalysis. In *Security, Forensics, Steganography and Watermarking of Multimedia Contents X*, volume 6819 of *Proc. SPIE*, 2008.

[48] A. D. Ker. A general framework for structural analysis of LSB replacement. In M. Barni, J. Herrera, S. Katzenbeisser, and F. Pérez-González, editors, *Information Hiding, 7th International Workshop*, volume 3727 of Lecture Notes in Computer Science, pages 296–311, Barcelona, Spain, June 6–8, 2005. Springer-Verlag, Berlin.

[49] A. D. Ker. Steganalysis of LSB matching in grayscale images. *IEEE Signal Processing Letters*, 12(6):441–444, June 2005.

[50] A. D. Ker. Batch steganography and pooled steganalysis. In J.-L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of Lecture Notes in Computer Science, pages 265–281, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.

[51] A. D. Ker. A capacity result for batch steganography. *IEEE Signal Processing Letters*, 14(8):525–528, 2007.

[52] A. D. Ker. The ultimate steganalysis benchmark? In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 141–148, Dallas, TX, September 20–21, 2007.

[53] A. D. Ker and R. Böhme. Revisiting weighted stego-image steganalysis. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, volume 6819, pages 5 1–5 17, San Jose, CA, January 27–31, 2008.

[54] A. D. Ker, T. Pevný, J. Kodovský, and J. Fridrich. The Square Root Law of steganographic capacity. In A. D. Ker, J. Dittmann, and J. Fridrich, editors, *Proceedings of the 10th ACM Multimedia & Security Workshop*, pages 107–116, Oxford, UK, September 22–23, 2008.

[55] Y. Kim, Z. Duric, and D. Richards. Modified matrix encoding technique for minimal distortion steganography. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of Lecture Notes in Computer Science, pages 314–327, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.

[56] J. Kodovský and J. Fridrich. On completeness of feature spaces in blind steganalysis. In A. D. Ker, J. Dittmann, and J. Fridrich, editors, *Proceedings of the 10th ACM Multimedia & Security Workshop*, pages 123–132, Oxford, UK, September 22–23, 2008.

[57] J. Kodovský and J. Fridrich. Calibration revisited. In J. Dittmann, S. Craver, and J. Fridrich, editors, *Proceedings of the 11th ACM Multimedia & Security Workshop*, pages 63–74, Princeton, NJ, September 7–8, 2009.

[58] J. Kodovský, T. Pevný, and J. Fridrich. Modern steganalysis can detect YASS. In N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XII*, volume 7541, pages 02–01–02–11, San Jose, CA, January 17–21, 2010.

[59] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Can be obtained from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

[60] L. Mevel and L. Finesso. Asymptotical statistics of misspecified hidden Markov models. *Automatic Control, IEEE Transactions on*, 49(7):1123–1132, July 2004.

[61] P. Moulin and Y. Wang. New results on steganographic capacity. In *Proceedings of the Conference on Information Sciences and Systems, CISS*, Princeton, NJ, March 17–19, 2004.

[62] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, September 25 1993.

[63] NRCS photo gallery. http://photogallery.nrcs.usda.gov/, accessed April 2004.

[64] T. Pevný, P. Bas, and J. Fridrich. Steganalysis by subtractive pixel adjacency matrix. In J. Dittmann, S. Craver, and J. Fridrich, editors, *Proceedings of the 11th ACM Multimedia & Security Workshop*, pages 75–84, Princeton, NJ, September 7–8, 2009.

[65] T. Pevný, T. Filler, and P. Bas. Using high-dimensional image models to perform highly undetectable steganography. In R. Böhme and R. Safavi-Naini, editors, *Information Hiding, 12th International Workshop*, volume 6387 of Lecture Notes in Computer Science, pages 161–177, Calgary, Canada, June 28–30, 2010. Springer-Verlag, New York.

[66] T. Pevný, T. Filler, and P. Bas. Using high-dimensional image models to perform highly undetectable steganography. In R. Böhme and R. Safavi-Naini, editors, *Information Hiding, 12th International Workshop*, Lecture Notes in Computer Science, Calgary, Canada, June 28–30, 2010. Springer-Verlag, New York.

[67] T. Pevný and J. Fridrich. Merging Markov and DCT features for multi-class JPEG steganalysis. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, pages 3 1–3 14, San Jose, CA, January 29–February 1, 2007.

[68] T. Pevný and J. Fridrich. Benchmarking for steganography. In K. Solanki, K. Sullivan, and U. Madhow, editors, *Information Hiding, 10th International Workshop*, volume 5284 of Lecture Notes in Computer Science, pages 251–267, Santa Barbara, CA, June 19–21, 2008. Springer-Verlag, New York.

[69] N. Provos. Defending against statistical steganalysis. In *10th USENIX Security Symposium*, pages 323–335, Washington, DC, August 13–17, 2001.

[70] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, January 2009.

[71] V. Sachnev, H. J. Kim, and R. Zhang. Less detectable JPEG steganography method based on heuristic optimization and BCH syndrome coding. In J. Dittmann, S. Craver, and J. Fridrich, editors, *Proceedings of the 11th ACM Multimedia & Security Workshop*, pages 131–140, Princeton, NJ, September 7–8, 2009.

[72] P. Sallee. Model-based methods for steganography and steganalysis. *International Journal of Image Graphics*, 5(1):167–190, 2005.

[73] A. Sarkar, L. Nataraj, B. S. Manjunath, and U. Madhow. Estimation of optimum coding redundancy and frequency domain analysis of attacks for YASS - a randomized block based hiding scheme. In *Proceedings IEEE, International Conference on Image Processing, ICIP 2008*, pages 1292–1295, San Diego, CA, October 12–15, 2008.

[74] A. Sarkar, K. Solanki, U. Madhow, and B. S. Manjunath. Secure steganography: Statistical restoration of the second order dependencies for improved security. In *Proceedings IEEE, International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–277–II–280, April 15–20, 2007.

[75] D. Schönfeld and A. Winkler. Embedding with syndrome coding based on BCH codes. In S. Voloshynovskiy, J. Dittmann, and J. Fridrich, editors, *Proceedings of the 8th ACM Multimedia & Security Workshop*, pages 214–223, Geneva, Switzerland, September 26–27, 2006.

[76] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, 4:142–163, 1959.

[77] Y. Q. Shi, C. Chen, and W. Chen. A Markov process based approach to effective attacking JPEG steganography. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of Lecture Notes in Computer Science, pages 249–264, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.

[78] V. Sidorenko and V. Zyablov. Decoding of convolutional codes using a syndrome trellis. *IEEE Transactions on Information Theory*, 40(5):1663–1666, 1994.

[79] M. Sidorov. Hidden Markov models and steganalysis. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 6th ACM Multimedia & Security Workshop*, pages 63–67, Magdeburg, Germany, September 20–21, 2004.

[80] K. Solanki, A. Sarkar, and B. S. Manjunath. YASS: Yet another steganographic scheme that resists blind steganalysis. In T. Furon, F. Cayre, G. Doërr, and P. Bas, editors, *Information Hiding, 9th International Workshop*, volume 4567 of Lecture Notes in Computer Science, pages 16–31, Saint Malo, France, June 11–13, 2007. Springer-Verlag, New York.

[81] K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran. Provably secure steganography: Achieving zero K–L divergence using statistical restoration. In *Proceedings IEEE, International Conference on Image Processing, ICIP 2006*, pages 125–128, Atlanta, GA, October 8–11, 2006.

[82] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[83] M. van Dijk and F. Willems. Embedding information in grayscale images. In *Proceedings of the 22nd Symposium on Information and Communication Theory*, pages 147–154, Enschede, The Netherlands, May 15–16, 2001.

[84] S. Verdú and T. Weissman. Erasure entropy. In *Proc. of ISIT*, Seattle, WA, July 9–14, 2006.

[85] S. Verdú and T. Weissman. The information lost in erasures. *IEEE Transactions on Information Theory*, 54(11):5030–5058, November 2008.

[86] A. Viterbi and J. Omura. Trellis encoding of memoryless discrete-time sources with a fidelity criterion. *IEEE Transactions on Information Theory*, 20(3):325–332, May 1974.

[87] F. Wang and D. P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64(5):056101, 2001. arXiv:cond-mat/0107006v1.

[88] Y. Wang and P. Moulin. Steganalysis of block-structured stegotext. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VI*, volume 5306, pages 477–488, San Jose, CA, January 19–22, 2004.

[89] Y. Wang and P. Moulin. Perfectly secure steganography: Capacity, error exponents, and code constructions. *IEEE Transactions on Information Theory, Special Issue on Security*, 55(6):2706–2722, June 2008.

[90] A. Westfeld. High capacity despite better steganalysis (F5 – a steganographic algorithm). In I. S. Moskowitz, editor, *Information Hiding, 4th International Workshop*, volume 2137 of Lecture Notes in Computer Science, pages 289–302, Pittsburgh, PA, April 25–27, 2001. Springer-Verlag, New York.

[91] A. Westfeld and R. Böhme. Exploiting preserved statistics for steganalysis. In J. Fridrich, editor, *Information Hiding, 6th International Workshop*, volume 3200 of Lecture Notes in Computer Science, pages 82–96, Toronto, Canada, May 23–25, 2004. Springer-Verlag, Berlin.

[92] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction (Stochastic Modelling and Applied Probability)*. Springer-Verlag, Berlin Heidelberg, 2nd edition, 2003.

[93] R. Zhang, V. Sachnev, and H. J. Kim. Fast BCH syndrome coding for steganography. In S. Katzenbeisser and A.-R. Sadeghi, editors, *Information Hiding, 11th International Workshop*, volume 5806 of Lecture Notes in Computer Science, pages 31–47, Darmstadt, Germany, June 7–10, 2009. Springer-Verlag, New York.

[94] W. Zhang, S. Wang, and X. Zhang. Improving embedding efficiency of covering codes for applications in steganography. *IEEE Communications Letters*, 11:680–682, August 2007.

[95] W. Zhang and X. Wang. Generalization of the ZZW embedding construction for steganography. *Information Forensics and Security, IEEE Transactions on*, 4(3):564–569, September 2009.

[96] W. Zhang, X. Zhang, and S. Wang. Maximizing steganographic embedding efficiency by combining Hamming codes and wet paper codes. In K. Solanki, K. Sullivan, and U. Madhow, editors, *Information Hiding, 10th International Workshop*, volume 5284 of Lecture Notes in Computer Science, pages 60–71, Santa Barbara, CA, June 19–21, 2008. Springer-Verlag, New York.

[97] W. Zhang and X. Zhu. Improving the embedding efficiency of wet paper codes by paper folding. *IEEE Signal Processing Letters*, 16(9):794–797, September 2009.

[98] X. Zhang, W. Zhang, and S. Wang. Efficient double-layered steganographic embedding. *Electronics Letters*, 43:482–483, April 2007.